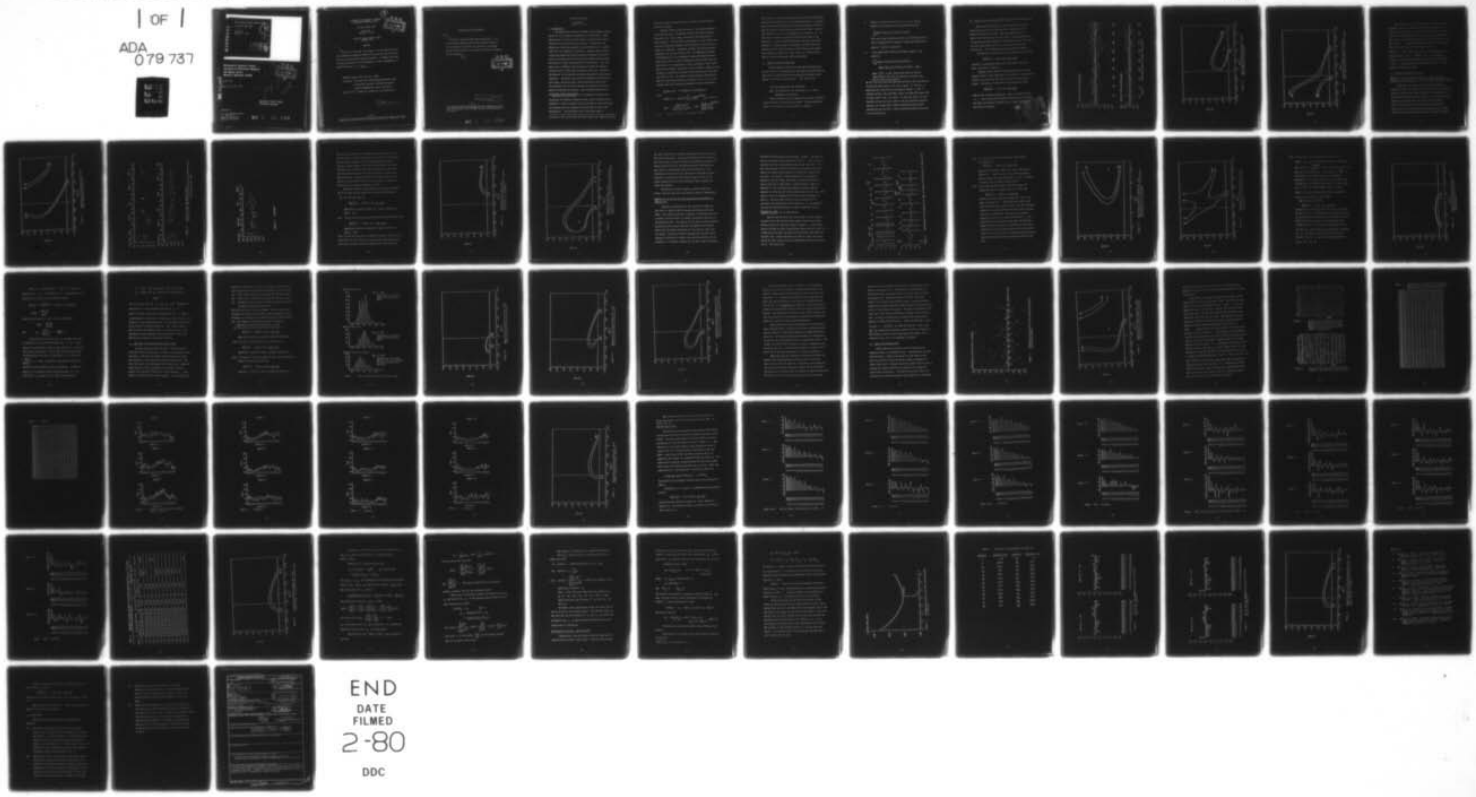
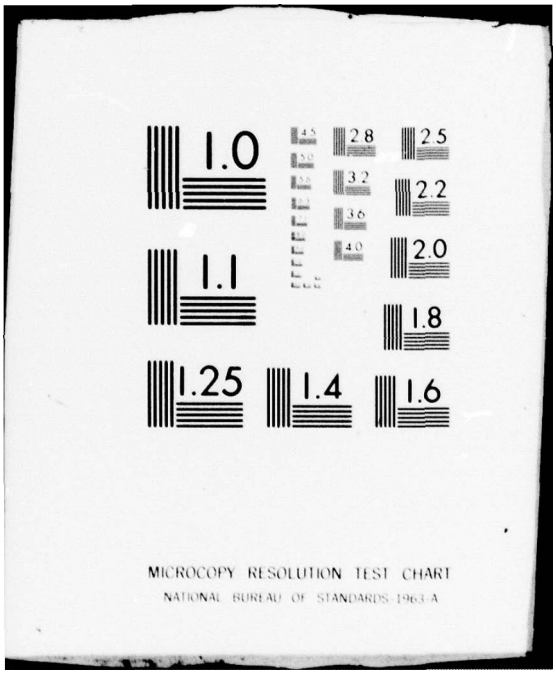
AD-A079 737 WISCONSIN UNIV-MADISON MATHEMATICS RESEARCH CENTER F/G 12/1
A STUDY OF REAL DATA.(U)
OCT 79 G CHEN, G E BOX DAAG29-75-C-0024
UNCLASSIFIED MRC-TSR-2002 NL

| OF |
ADA
079 737



END
DATE
FILMED
2-80
DDC


1.0
1.1
1.25
1.4
1.6
2.0
2.2
2.5
2.8
3.2
3.6
4.0
4.5
5.0
MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

MRC Technical Summary Report #2002

A STUDY OF REAL DATA

Gina Chen
George E. P. Box

LEVEL II

Mathematics Research Center
University of Wisconsin—Madison
610 Walnut Street
Madison, Wisconsin 53706

October 1979

Received July 27, 1979

Approved for public release
Distribution unlimited

Sponsored by

U. S. Army Research Office
P. O. Box 12211
Research Triangle Park
North Carolina 27709

UNIVERSITY OF WISCONSIN - MADISON
MATHEMATICS RESEARCH CENTER

A STUDY OF REAL DATA

Gina Chen
George E. P. Box

Technical Summary Report #2002
October 1979

ABSTRACT

Nine sets of real data are analyzed. The distributions within the contaminated exponential power family which best describe these data sets are obtained by maximum likelihood. It appears that heavy tailed distributions are often produced by secular inhomogeneity in mean and variance. → to page - B -

AMS(MOS) Subject Classification: 62G35

Key Words: real data sets, contaminated exponential power distribution, maximum likelihood estimation, secular inhomogeneity, serial correlation

Work Unit #4 - Probability, Statistics, and Combinatorics

Sponsored by the United States Army under Contract No. DAAG-29-75-C-0024.

SIGNIFICANCE AND EXPLANATION

cont.

A specific robust estimator would behave well if, in the real world, data occurred of the kind which favored it. In this paper more classical sets of data are considered and their characteristics are studied in relation to proposed robust estimators.

The responsibility for the wording and views expressed in this descriptive summary lies with MRC, and not with the authors of this report.

A STUDY OF REAL DATA

Gina Chen
George E. P. Box

1. Introduction.

The appropriate procedure to adopt in data analysis depends heavily on how the data are supposed to be distributed. Exact distributions of data are rarely known since errors are usually contributed to by many different factors. The central limit theorem has often been appealed to by statisticians as a justification for the assumption of normality. However, more recently it has been argued that heavier-tailed distributions and distributions containing outliers might be more likely than the normal distribution to describe the behavior of real errors. New estimators have, therefore, been proposed which would be expected to behave well if these other assumptions were true. Studies of the behavior of the proposed estimators have usually been made by simulation. Typically, some feared distribution or distributions were chosen from which pseudo random samples have been drawn. The estimators have then been computed and various of their properties calculated and compared. Most frequently, efficiencies $\left(\frac{1/(n \text{ times Fisher information})}{\text{variance of the estimator}}\right)$ (or asymptotic variances) were calculated for different estimators under various distribution assumptions, a robust estimator would be one having a high efficiency (or comparatively low asymptotic variance) over all the tested distributions. The difficulty here is that such study is

---

Sponsored by the United States Army under Contract No. DAAG-29-75-C-0024.

necessarily rather subjective since it depends on what distributions are included.

Stigler (1977) raised the question "Do robust estimators work with real data?" He applied different proposed estimators to several sets of real data, and measured the performance of each estimator. With the data he considered, the 10% trimmed mean was one of the best estimators and the ordinary sample mean came close as a competitor. Estimators selected from the best modern contenders did not come out very well. As commented by Stigler "the alternatives to an independent identically normally distributed sample that have been considered by modern workers are too restricted or too exaggerated to reflect accurately 'real' data." Later in this chapter, therefore, we shall examine a number of sets of data as a contribution towards finding out what data from the real world is like. As a means of approximately characterizing data sets and relating them to the results of TSR #1997, we shall consider again the contaminated exponential power distribution

$$P_c(y|\theta,\sigma,\beta,\alpha) = (1-\alpha)P(y|\theta,\sigma,\beta)+\alpha P(y|\theta,k\sigma,\beta)$$

with

$$P(y|\theta,\sigma,\beta) = w(\beta)\sigma^{-1}\exp\left[-c(\beta)\left|\frac{y-\theta}{\sigma}\right|^{2/(1+\beta)}\right] \quad -\infty < y < \infty$$

$$w(\beta) = \frac{\{\Gamma[\frac{3}{2}(1+\beta)]\}^{1/2}}{(1+\beta)\{\Gamma(\frac{1}{2}(1+\beta)]\}^{3/2}} \qquad c(\beta) = \left\{\frac{\Gamma[\frac{3}{2}(1+\beta)]}{\Gamma[\frac{1}{2}(1+\beta)]}\right\}^{1/(1+\beta)}$$

and $\sigma > 0$, $-\infty < \theta < \infty$, $-1 < \beta \leq 1$, $k > 1$.

This provides a reasonably broad family of symmetric distributions containing both light-tailed and heavy-tailed members. To characterize the parent distribution for each set of data we shall fit this distribution to the data by maximimum likelihood. There are four parameters θ, σ, α, β altogether (k is fixed at 3). In this chapter, however, we are particularly interested in two parameters, α and β, which characterize the shape of the distribution. For each data set, therefore, maximum likelihood estimates for all four parameters will be obtained and approximate confidence regions for (β,α) with θ,σ fixed at their maximum likelihood estimates will be calculated.

2. Analyzing the Real Data Sets

In this study of searching for the parent distributions of real data sets, one must keep in mind that we are assuming the data set is a random sample from some particular parent distribution in the family $P_c(y|\theta,\sigma,\beta,\alpha)$. This implies that

(i) The observations are independent.

(ii) The observations are stationary, i.e., have a fixed mean and variance.

These assumptions must be checked, because their violation could invalidate the estimation procedure. If the assumptions appear to be satisfied, one can proceed as follows:

1. Maximum likelihood estimates for θ, σ, β, α may be obtained by maximizing the log likelihood function,

$$\sum_{i=1}^{n} \log[(1-\alpha)P(y_i|\theta,\sigma,\beta)+\alpha P(y_i|\theta,3\sigma,\beta)].$$

This may be done by evaluating this log likelihood function over a suitable grid of parameter values and obtaining the maximum $(\hat{\theta},\hat{\sigma},\hat{\beta},\hat{\alpha})$ numerically.

2. A very approximate $100(1-\epsilon)\%$ confidence region is now given by

$$-2\sum_{i=1}^{n}\{\log[(1-\alpha)P(y_i|\hat{\theta},\hat{\sigma},\beta)+\alpha P(y_i|\hat{\theta},3\hat{\sigma},\beta)]$$
$$-\log[(1-\hat{\alpha})P(y_i|\hat{\theta},\hat{\sigma},\hat{\beta})+\hat{\alpha}P(y_i|\hat{\theta},3\hat{\sigma},\hat{\beta})]\} < \chi^2_\epsilon(2)$$

where $\chi^2_\epsilon(2)$ is the 100ϵ percent point of the chi-square distribution with two degrees of freedom since we have fixed two parameters.

We will thus have some idea about what positions the distributions of real data might take in the (β,α) space. In TSR#1997, we mentioned that trade offs can be expected between α and β yielding oblique ridgy confidence regions oriented from upper left to lower right in the β,α plane. It turns out that it is also possible to have two local maxima in the log likelihood function surface, one representing a highly contaminated short-tailed distribution and the other a heavy-tailed distribution with small or no contamination.

3. Daily Changes in Price of IBM Common Stock (Box and Jenkins)

Box and Jenkins (1970) give records of daily IBM common stock closing prices from 17th May 1961 - 2nd November 1962 (Series B) and 29th June 1959 - 30th June 1960 (Series B'). Series B has 369 observations and Series B' has 255 observations. It has been suggested (Bachelier, 1900) that the first differences of stock prices behave like i.i.d. random sample from some distribution. The differenced series are plotted in Figure 1.

For Series B, considering first the whole sample, we have,

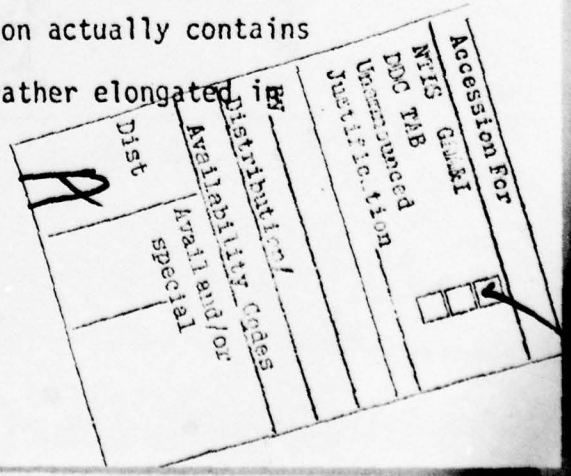
$$(\hat{\theta},\hat{\sigma},\hat{\beta},\hat{\alpha}) = (0.0,\ 6.65,\ \underline{0.60},\ \underline{0.03})$$

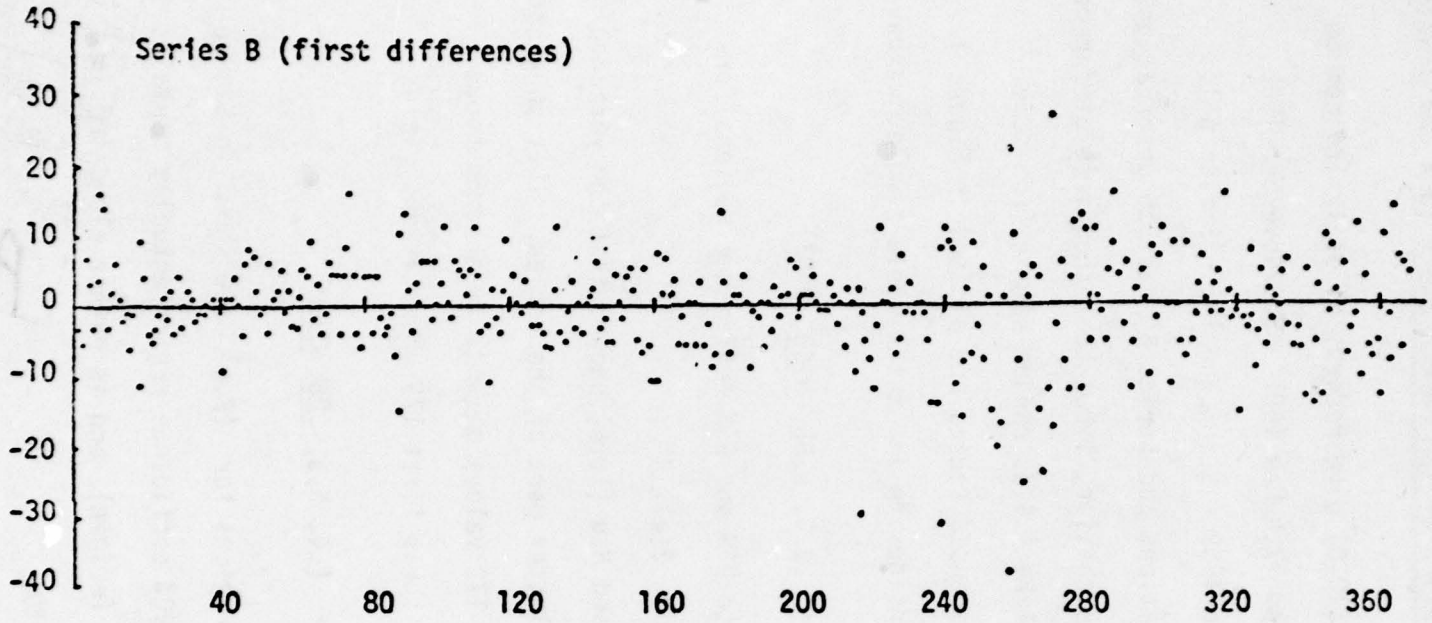
and both very approximate 90% and 95% confidence regions for (β,α) are given in Figure 2(a).

Wichern, Miller and Hsu (1976) showed that the variance became unstable in the latter part of the series. They suggested, however, that the first 179 values seem to be a homogeneous sample. An analysis of these first 179 observations yields

$$(\hat{\theta},\hat{\sigma},\hat{\beta},\hat{\alpha}) = (.4,\ 5.3,\ \underline{.25},\ \underline{0.00})\ .$$

90% and 95% confidence regions for (β,α) are shown in Figure 2(b). This time the 90% confidence region actually contains the normal distribution (origin), and is rather elongated in the direction expected.





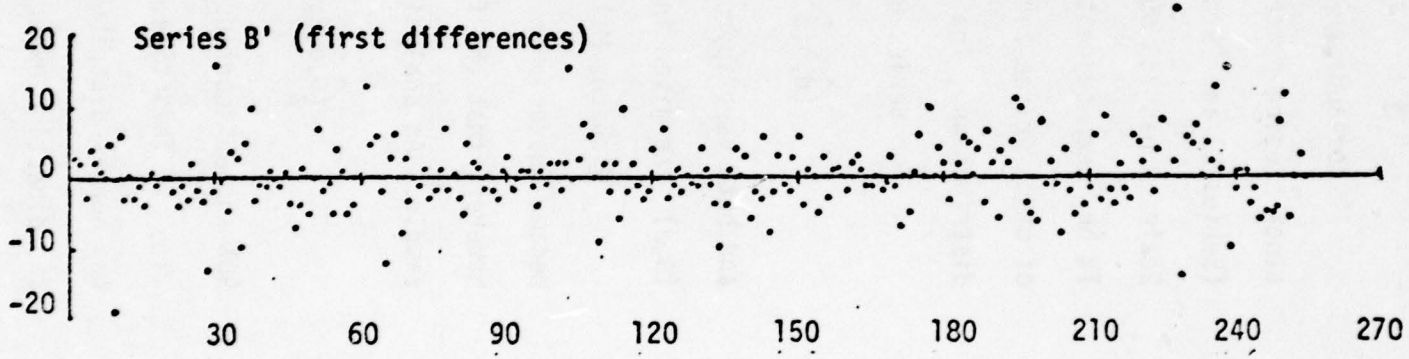


Figure 1. First Differences for Series B and Series B'

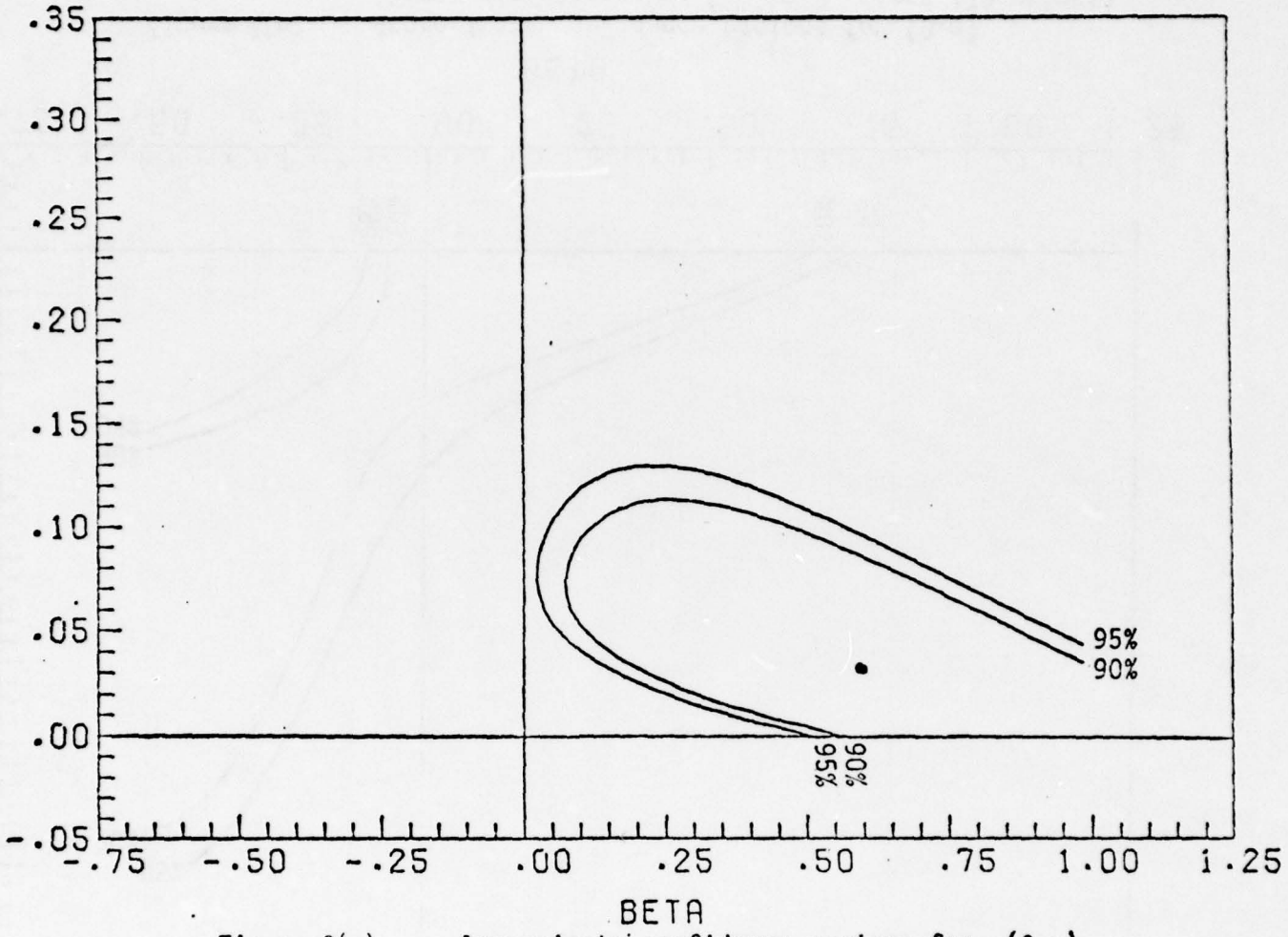


Figure 2(a). Approximate confidence regions for (β,α) (First differences of Series B)

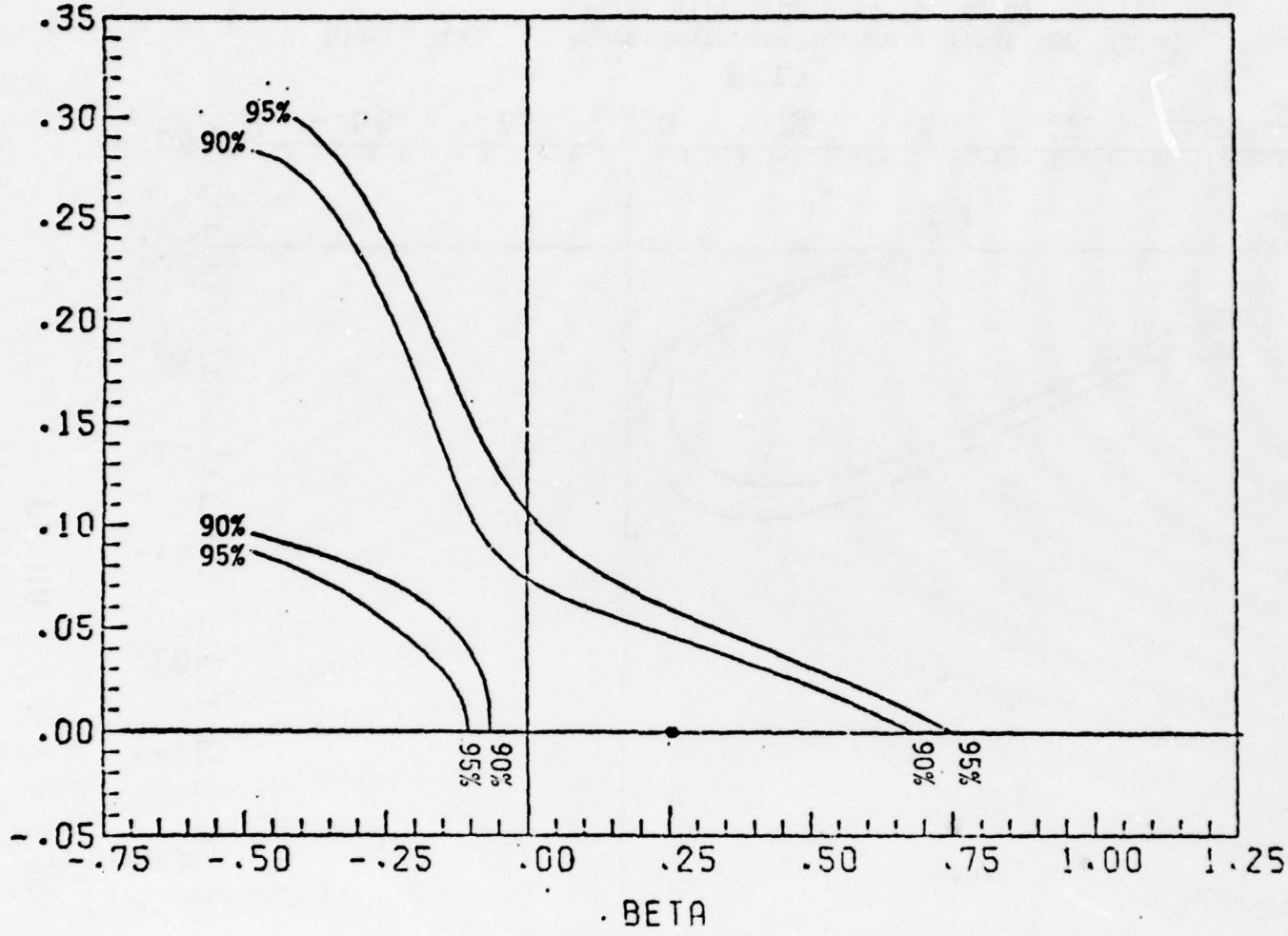


Figure 2(b). Approximate confidence regions for (β,α)
(First differences of Series B, first 179 values)

The example shows that for these data, tail heaviness of the whole sample arises because of changes in variance in the latter part of the series. It seems that for the first 179 observations the parent distribution could very well be represented by a normal, or slightly contaminated, normal distribution. As for Series B', maximum likelihood estimation was carried out for the first differences of the series to obtain $(\hat{\theta},\hat{\sigma},\hat{\beta},\hat{\alpha})$ = (0.0, 3.54, 0.20, 0.16) and the approximate confidence regions shown in Figure 2(c). The confidence regions are elongated as expected showing the trade off between β and α. A contaminated normal distribution with $\alpha \simeq .25$ could well represent the data.

4. Measures of Velocity of Light

Measurement of the Velocity of Light in a Partial Vacumm by Michelson, Pease and Pearson (Michelson, Pease and Pearson (1935))

During the period September 1929 to March 1933, the velocity of light was measured at the Irvine Ranch near Santa Ana, California by Michelson, Pease and Pearson. The observations were made by the rotating-mirror method. In all 233 series were recorded, the averages for which are plotted in Figure 3. Every series represents two to eight sets of observations taken during over about an hour's time with each set of observations usually furnishing three values of velocity of light.

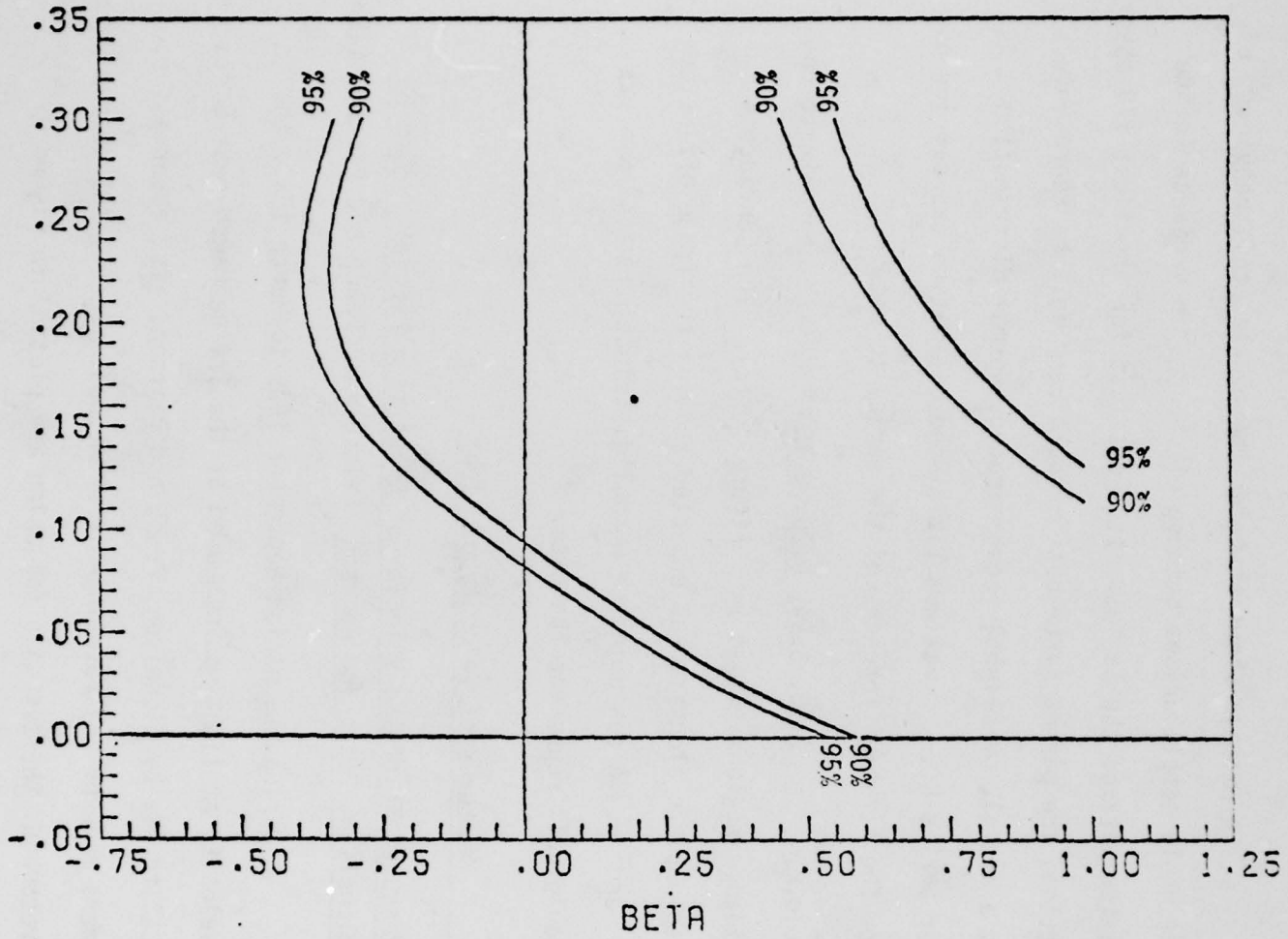


Figure 2(c). Approximate confidence regions for (β,α)
(First differences of Series B')

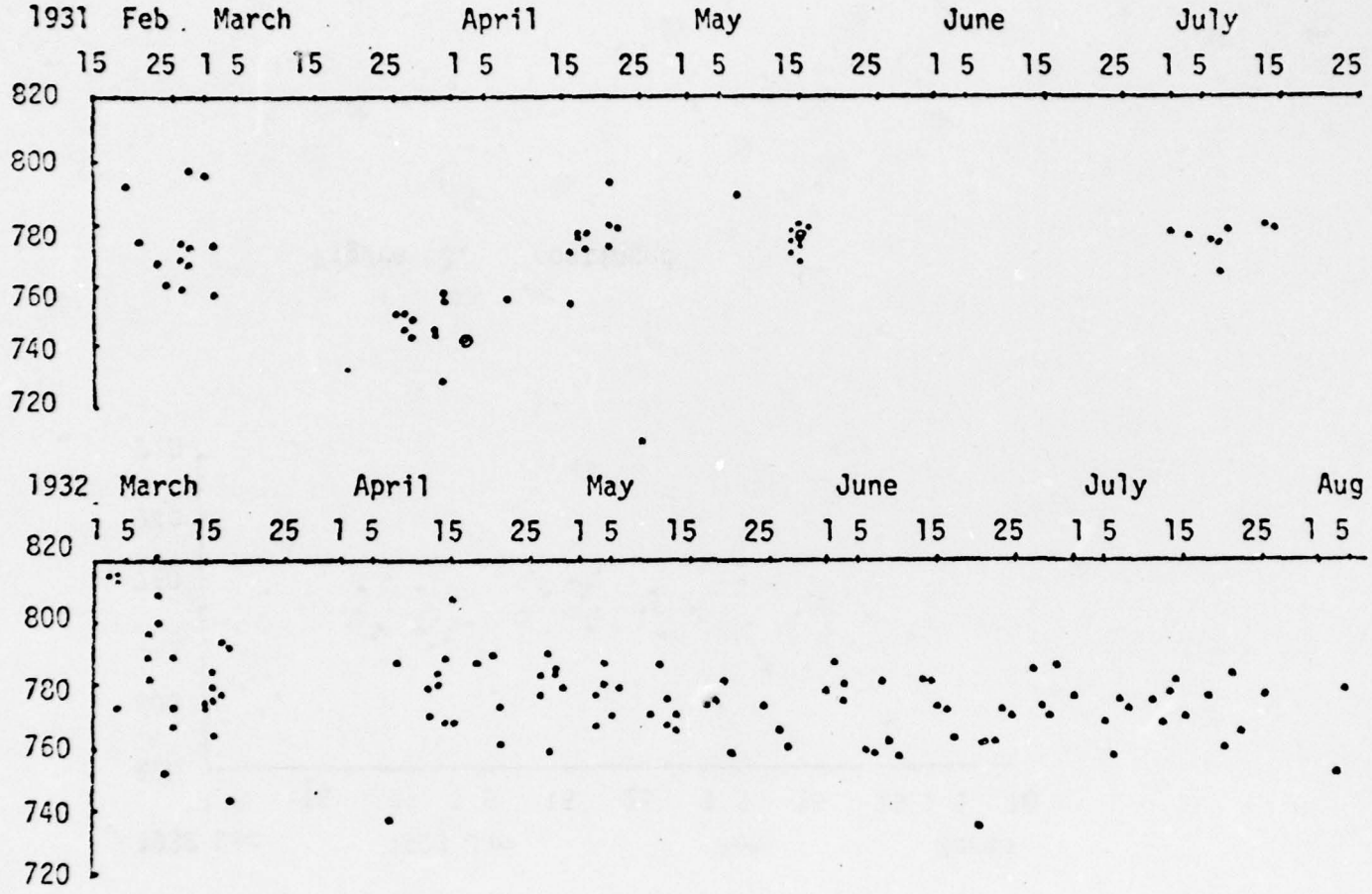


Figure 3. Michelson, Pease and Pearson's data
(Mean velocity - 299.000) measured between Feb.1931-March 1933

(continued on next page)

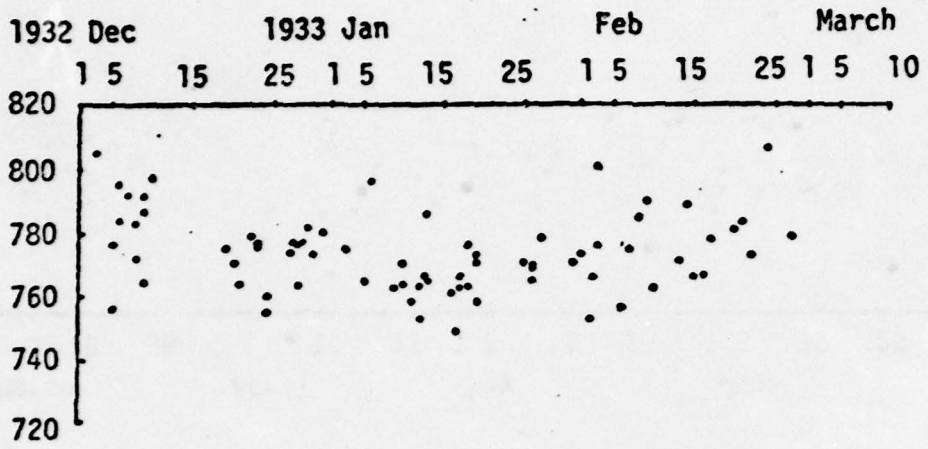


Figure 3. continued

All these observations were actually made between February 1931 and March 1933. In 1931, observations were not made continuously over the year; instead, a group of observations were made in February, another group in late March and early April and so on. The means and variances of these groups fluctuate considerably. The 155 observations made after April 1932 were, however, much more homogeneous. Behavior of this kind is common where a new method of scientific measurement takes some time to "debug" and to settle down to produce homogeneous results.

We shall, therefore, again perform two analyses—the first for all the data and the second for the last 155 observations.

(i) For the whole data set

$$(\hat{\theta},\hat{\sigma},\hat{\beta},\hat{\alpha}) = (774.0,\ 14.6,\ \underline{.80},\ \underline{0.0})$$

Approximate confidence regions for (β,α) are shown in Figure 4(a).

(ii) For the 155 observations between April 1932 and February 1933

$$(\hat{\theta},\hat{\sigma},\hat{\beta},\hat{\alpha}) = (773.1,\ 9.47,\ \underline{-.35},\ \underline{0.10})$$

Approximate confidence regions for (β,α) are drawn in Figure 4(b).

Again, notice that the maximum likelihood estimates $(\hat{\beta},\hat{\alpha})$ move from (.80, 0.0) to (-.35, 0.10) as we leave out the inhomogeneous part of the data, and the confidence regions, which originally cover

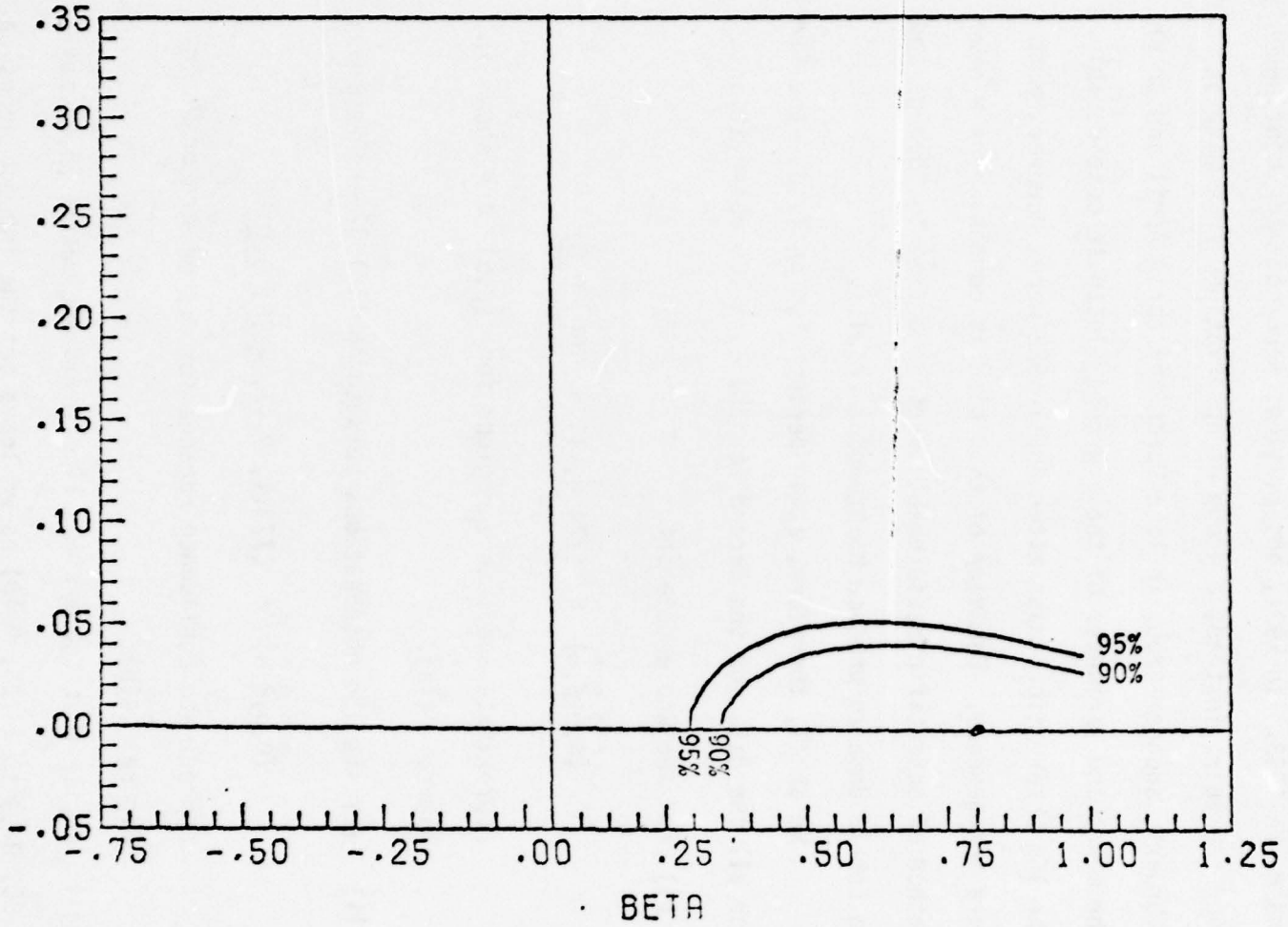


Figure 4(a). Confidence regions for (β,α)
(Measurements of the velocity of light by Michelson, Pease, Pearson)

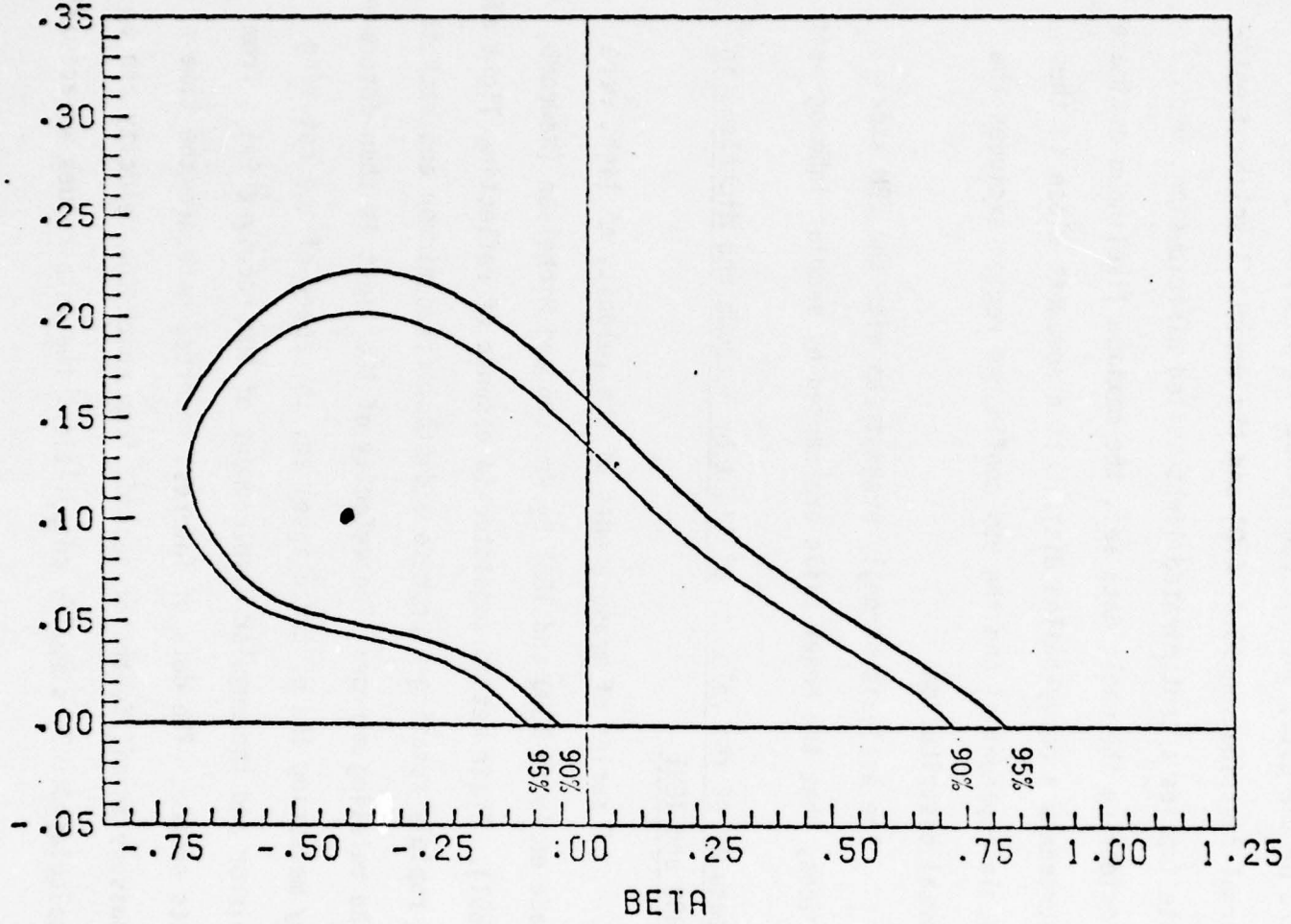


Figure 4(b). Confidence regions for (β,α)
(155 measurements of the velocity of light made between April 1932 and February 1933 by Michelson, Pease and Pearson)

the lower right corner, now cover the central part and some upper left part of the plane. Also note that while, for the homogeneous part of the data, the 90% confidence region actually contains the normal distribution (0.0, 0.0) and the maximum likelihood estimate implies a contaminated light-tailed distribution, when considering the whole data set, the maximum likelihood estimate represents a heavy-tailed distribution somewhat close to the double exponential and the 95% confidence region excludes the normal distribution.

The analysis strongly suggests, as with the IBM stock prices, that the heavy tails are caused by secular inhomogeneity.

Measures of the velocity of light by Newcomb and Michelson in 1880 and 1881

A series of measurements of the velocity of light were made early in 1880 and 1881 by Newcomb and Michelson (Newcomb, 1891). Their method consisted in essence of reflecting light off a rapidly rotating mirror to a distant fixed mirror and back to the rotating mirror. The velocity of the light is then determined by measuring the distance involved, the speed of the rotating mirror and the angular displacement of the received image from its source. The data of interest recorded here are the time of passage in millionths of second, from which the velocity can be calculated. To simplify computation, the data values we actually

considered are (recorded time of passage - 17,000). The observations made on each day are plotted in Figure 5. We can see from the plot that there are appreciable day to day variations. This was to be expected because of factors such as weather conditions. Some of the factors which might have affected the results were recorded. For example, it was noticed that it was cloudy on August 25, and that images were faint for some observations so that on that day we might expect a larger variance. Again on September 15, the instrument was adjusted, and this might be expected to influence both the mean and variance on that day. With these in mind, any assumption of constant mean and variance is dubious. One way to get rid of day to day variation is to standardize the observations within each day, i.e., to calculate the mean and standard deviation for data on each day, and use $\frac{\text{observation - mean}}{\text{standard deviation}}$ as new observations.

What we have done with this set of data is, first, analyze the whole set of the original data; second, leave out the three obvious outliers as shown by an arrow in Figure 5 ; and finally, analyze the data set after standardizing within each day. Note that in standardizing the data, the three obvious outliers were discarded, and when only a single observation was made in a day, this observarion was also excluded since no standard deviation was available for it. The results are

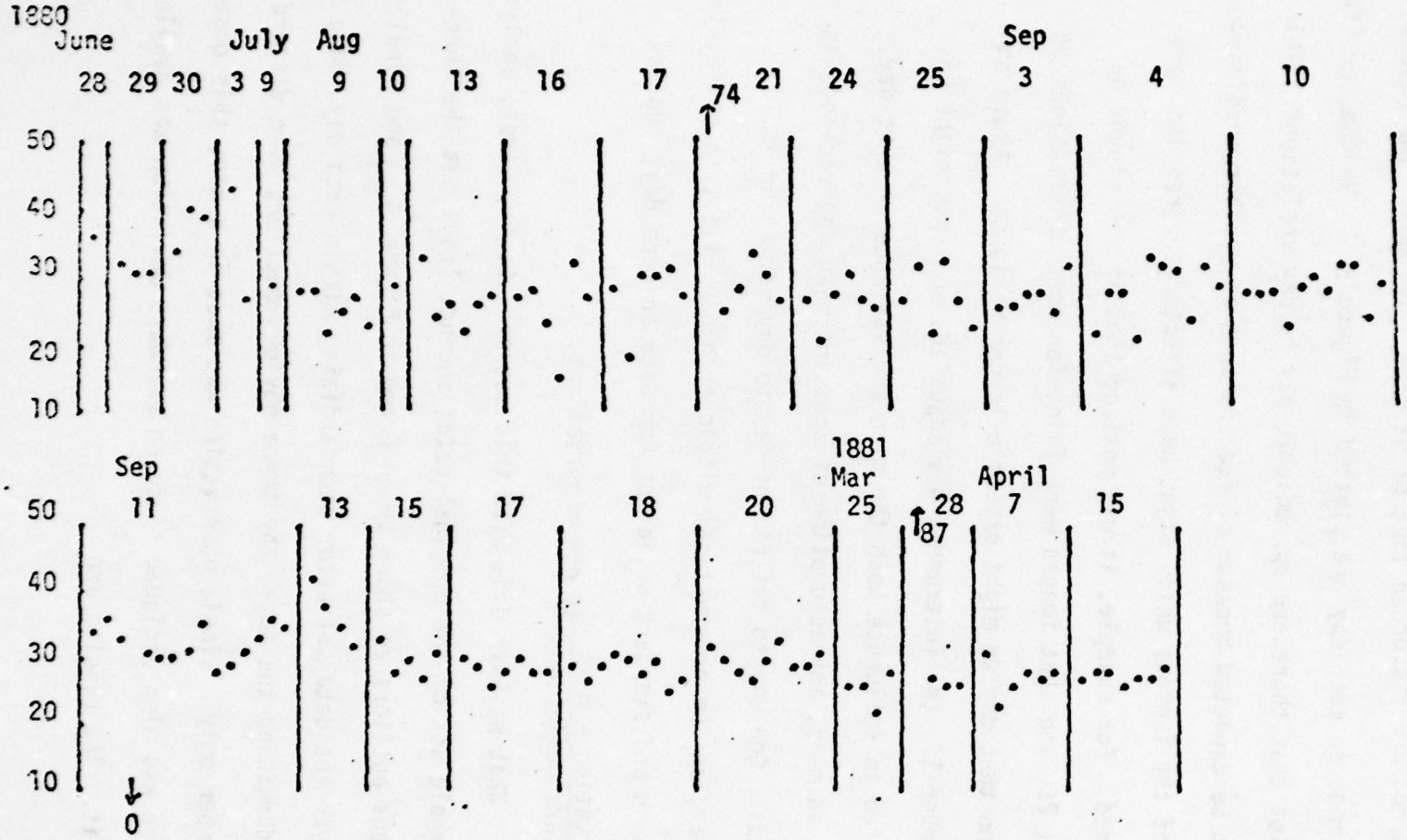


Figure 5. Measurements of the velocity of light by Newcomb and Michelson in 1880 and 1881.

(i) If we take the whole set of data as a random sample ($n = 150$), then

$$(\hat{\theta},\hat{\sigma},\hat{\beta},\hat{\alpha}) = (28.0,\ 4.16,\ \underline{1.0},\ \underline{.11})$$

Approximate confidence regions for (β,α) are given in Figure 6(a), which show that the parent distribution have very heavy tail (with both large β and α values) if the data set is really a random sample.

(ii) If we throw out three obvious wild observations and regard the rest as a random sample ($n = 147$), then

$$(\hat{\theta},\hat{\sigma},\hat{\beta},\hat{\alpha}) = (28.0,\ 4.3,\ \underline{1.0},\ \underline{0.0})$$

and the approximate confidence regions for (β,α) are given in Figure 6(b). These confidence regions have quite different shapes from the others, the 90% confidence region has two parts, located at the lower right and upper left plane, respectively. This came as no surprise because we know there is compensation between β and α. A distribution with heavy tail and small contamination could be somewhat similar to a distribution with light or moderate tail but high contamination, and for this particular set of data, both areas in the 90% confidence region contain distributions which can adequately describe the data.

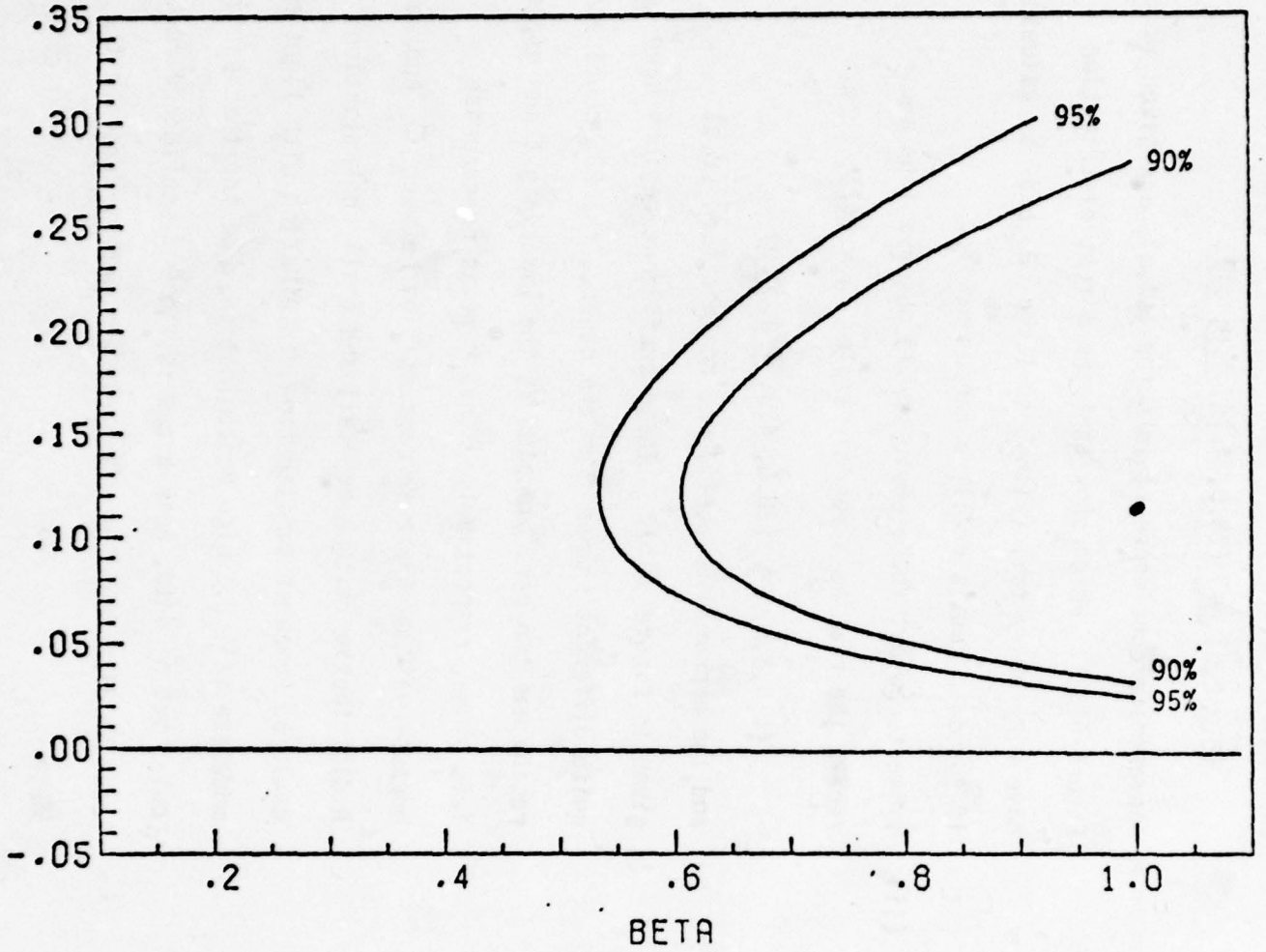


Figure 6(a) Approximate confidence regions for (β,α) (Measurements of the velocity of light by Newcomb and Michelson in 1880 and 1881)

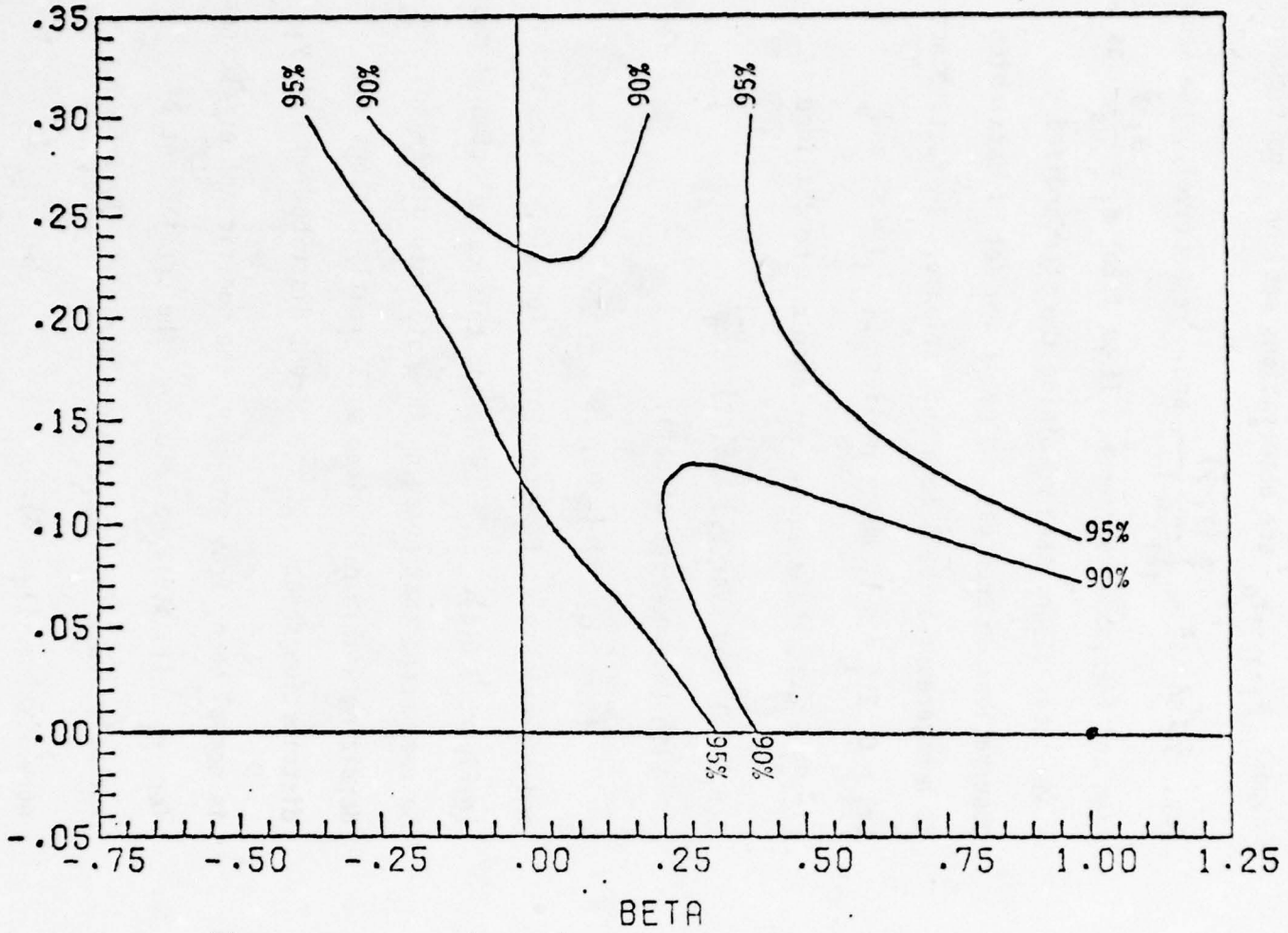


Figure 6(b) Approximate confidence regions for (β,α) (Newcomb and Michelson data after throwing out 3 outliers)

(iii) We standardize the observations made on each day.

Suppose $y_1,\ldots,y_n$ are observations made on a particular day, $\bar{y}$ and $s^2 = \sum_{i=1}^{n} \frac{(y_i-\bar{y})}{n-1}$ are, respectively, the sample mean and the sample variance. If we take $r_i = \frac{y_i-\bar{y}}{s}$ as a new observation, then combining the standardized observations on each day, we get a new set of data which is homogeneous in both mean and variance. In fact, since $\Sigma r_i = 0$ $\Sigma r_i^2 = n-1$, which implies $E(\Sigma r_i) = 0$ and $E(\Sigma r_i^2) = n-1$, the mean and variance can be obtained as $E(r_i) = 0$ and $Var(r_i) = E(r_i^2) = \frac{n-1}{n}$.

With this new set of data,

$$(\hat{\theta},\hat{\sigma},\hat{\beta},\hat{\alpha}) = (-.05, .94, \underline{-.65}, \underline{0.0})$$

and approximate confidence regions for (β,α) are shown in Figure 3.6(c). In considering this calculation it must be remembered that breaking the data into pieces and standardizing within each piece will greatly distort the distribution shape. If the parent distribution for y_i is normal then, from symmetry, the coefficient of skewness for r_i will still be zero, but the coefficient of kurtosis (λ_4) for r_i can deviate markedly from its normal value $(\lambda_4 = 0)$.

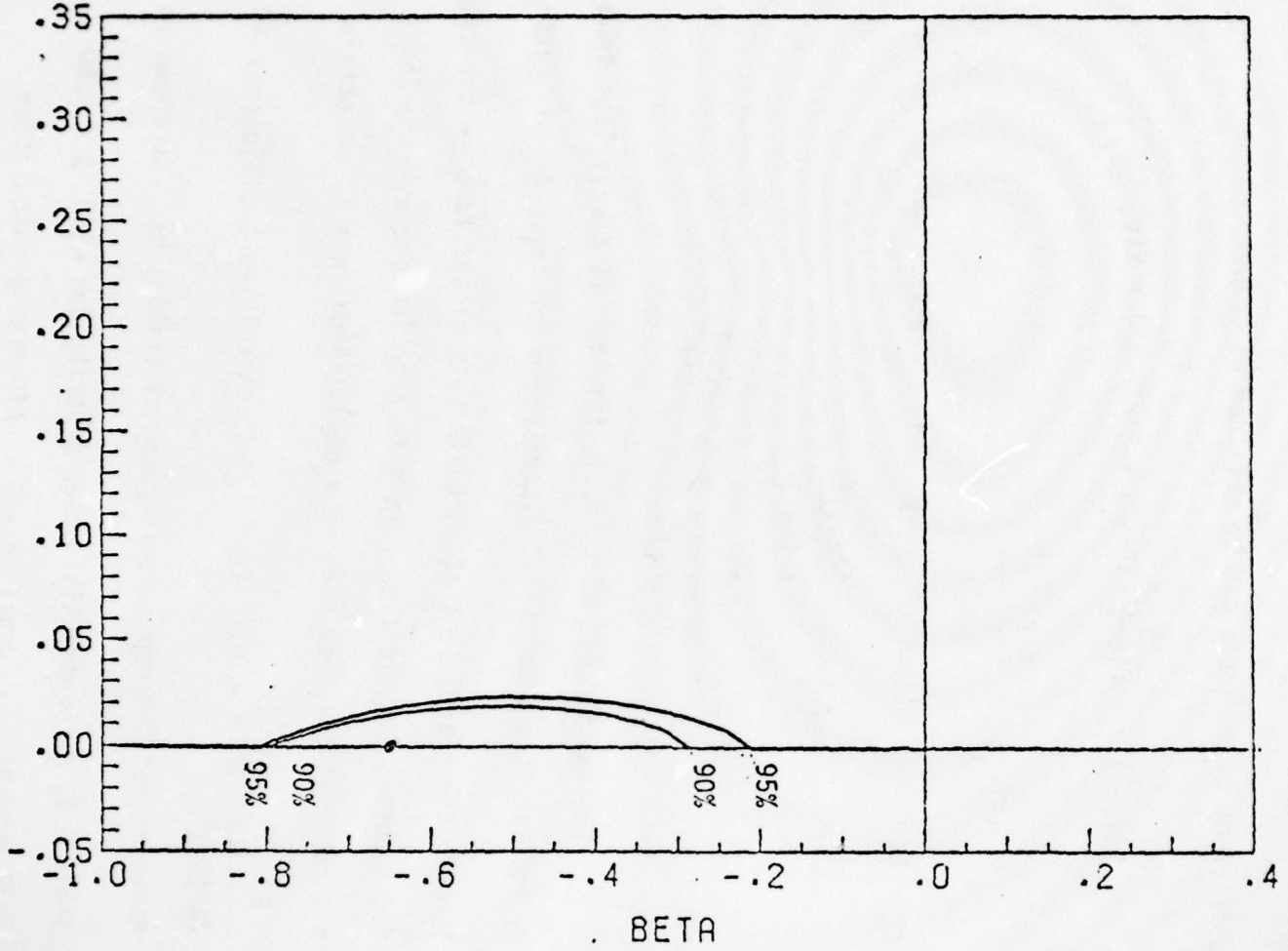


Figure 6(c) Approximate confidence regions for (β,α) (Newcomb and Michelson data after standardization)

Since r_i is independent of σ and s is a sufficient statistic for σ, r_i is independent of s. It follows that the coefficient of kurtosis can be obtained as below:

$$E(y_i-\bar{y})^4 = E((\frac{y_i-\bar{y}}{s})^4 \cdot s^4) = E(r_i^4 \cdot s^4) = E(r_i^4)E(s^4)$$

$$E(r_i^4) = \frac{E(y_i-\bar{y})^4}{E(s^4)}$$

From the distributions of $y_i-\bar{y}$ and s, we then have

$$E(r_i^4) = \frac{3(n-1)^3}{n\,(n+1)}$$

and $$\lambda_4 = \frac{E(r_i^4)}{Var(r_i)^2} - 3 = \frac{3(n-1)}{n+1} - 3.$$

The coefficient of kurtosis for r_i is always less than 0, indicating a flat tailed distribution for r_i. As n gets larger, the coefficient of kurtosis gets closer to 0 — the value for a normal distribution. In this case (on the average about six observations were made in a day), taking $n = 6$, we obtain $\frac{E(r_i)^4}{Var(r_i)^2} - 3 = -.858$. Lund (1967) calculated coefficients of kurtosis for the exponential power distributions. As shown in Table 3.1, an exponential power distribution with $\beta = -.50$ will produce λ_4 equal to -.81. After some calculation

β	-1.0	- .75	-.50	-.25	0.0	.25	.50	.75	1.00
λ_4	-1.20	-1.07	-.81	-.45	0.00	.55	1.22	2.03	3.00

Table 1

one can also show that for $\beta = -.55$, $\lambda_4 = -.87$. Therefore, we concluded that if the parent distribution for y_i is normal, one would expect that the distribution for r_i (when n is approximately 6) would have tail behavior similar to that of an exponential power distribution with $\beta = -.55$ which is very close to the maximum likelihood estimate of -.65. Thus it appears that the normal distribution alone, or maybe with some slight contamination, could adequately describe the data if there had not been daily changes in levels and variances.

5. Some Data Sets Investigated by Tocher (1928)

In 1928, Tocher published the paper "An Investigation of the Milk Yield of Dairy Cows," in which a large amount of data were given. These are the records about milk yields of dairy cows supplied by "The Scottish Milk Records Association." Various "characters" are investigated for each cow, e.g., estimated total quantity of milk in gallons for the entire lactation period, total quantity of butter fat, age of the cow, ..., etc. Tocher's investigation had several purposes. In one interesting

analysis he fitted Pearson curves to a number of large data sets. Three of these data sets which we will consider in our study are

(1) Total yield of butter fat in pounds for 449 cows at age 7.

(2) Total yield of butter fat in pounds for 251 cows at age 9.

(3) Quantity of milk in gallons for 240 cows with percentages of butter fat between 3.00 and 3.25.

The exact data values are not available. Tocher's paper gives frequency distributions for the data, with a grouping interval of 50 gallons for quantity of milk and 20 pounds for yield of butter fat. The frequency distributions are shown in Figure 7.

Following the procedure stated before, we have

(i) For total yield of butter fat for cows at age 7

$$(\hat{\theta},\hat{\sigma},\hat{\beta},\hat{\alpha}) = (249.75,\ 52.15,\ \underline{-.20},\ \underline{.01})$$

Approximate confidence regions are shown in Figure 8(a).

(ii) For total yield of butter fat for cows at age 9

$$(\hat{\theta},\hat{\sigma},\hat{\beta},\hat{\alpha}) = (265.8,\ 60.15,\ \underline{0.35},\ \underline{0.0})$$

Approximate confidence regions are given in Figure 8(b).

(iii) For quantity of milk in gallons for cows with percentages of butter fat between 3.00 and 3.25.

$$(\hat{\theta},\hat{\sigma},\hat{\beta},\hat{\alpha}) = (664.0,\ 176.0,\ \underline{0.25},\ \underline{0.0})$$

Approximate confidence regions are given in Figure 8(c).

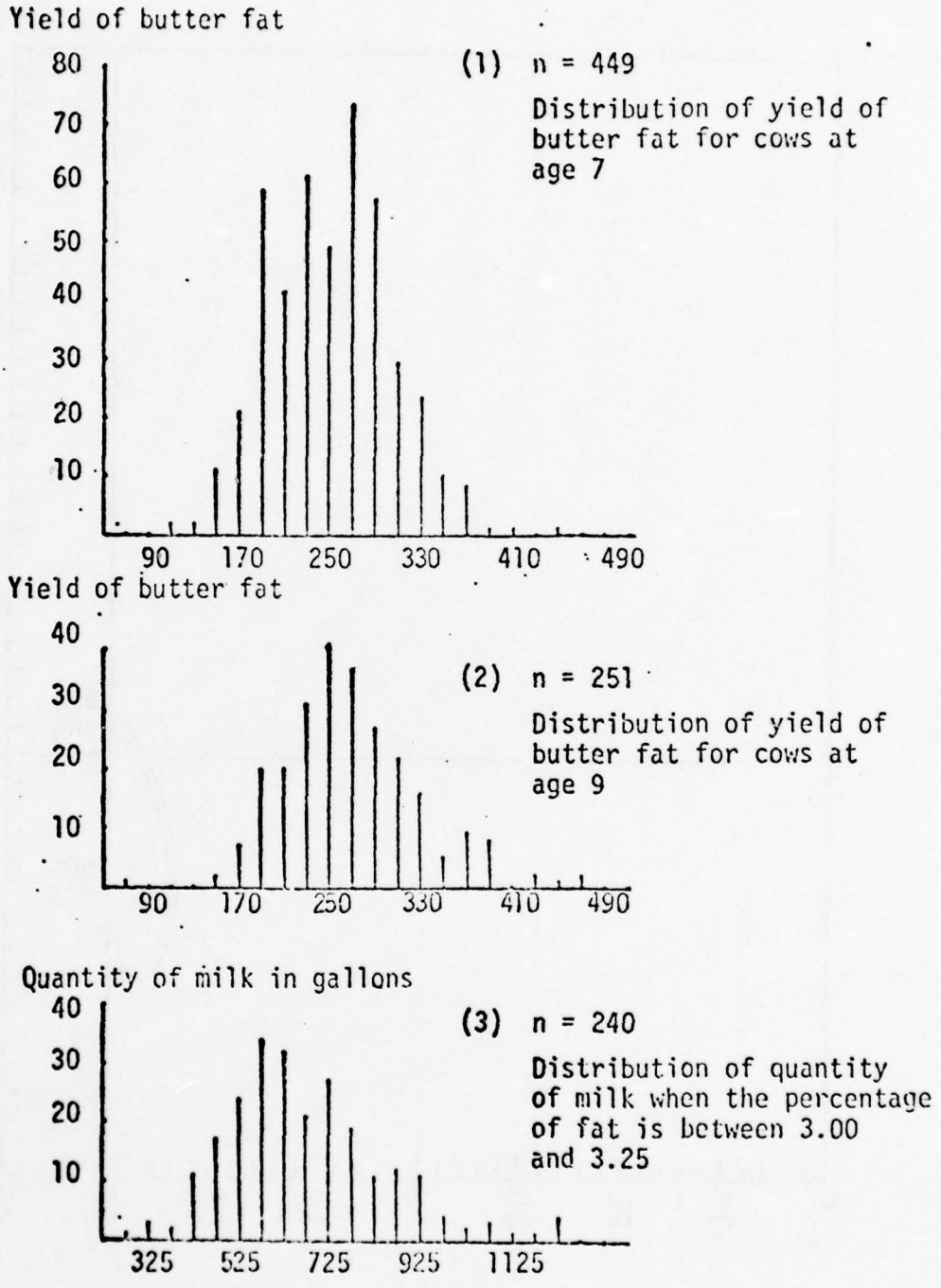


Figure 7 Frequency distributions for Tocher's data.

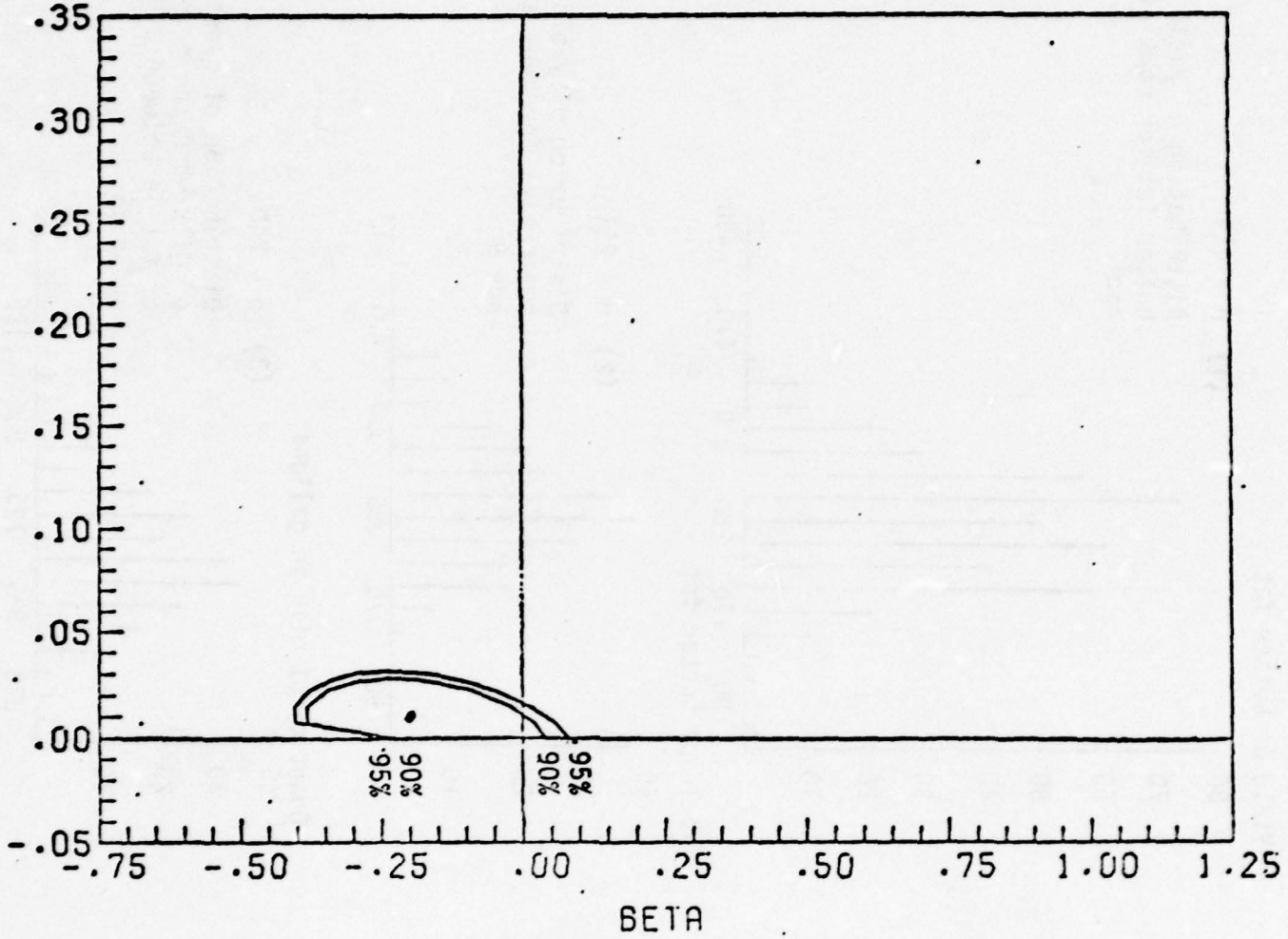


Figure 8(a) Approximate confidence regions for (β,α)
(Yield of butter fat for cows at age 7)

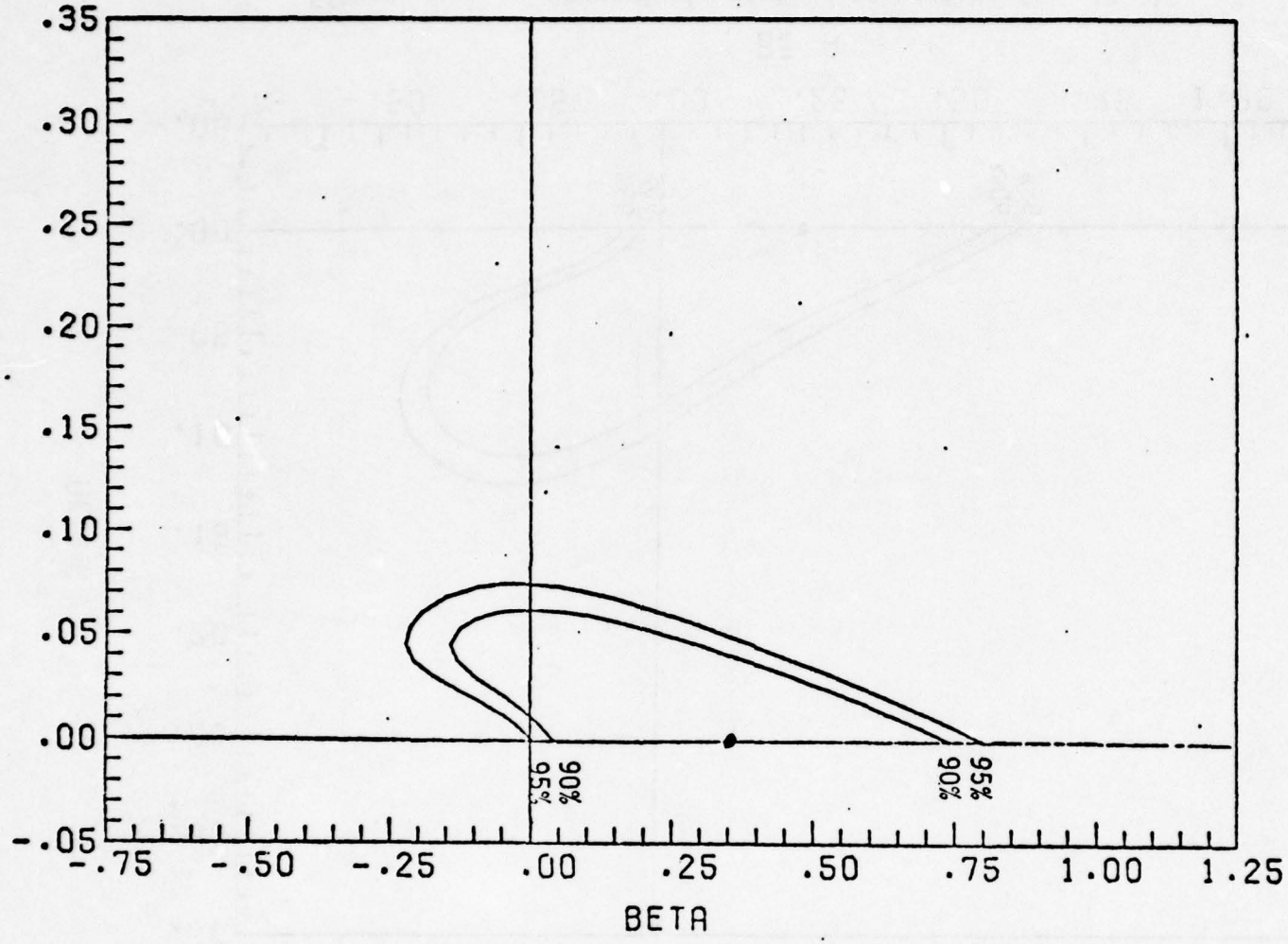


Figure 8(b) Approximate confidence regions for (β,α)
(Yield of butter fat for cows at age 9)

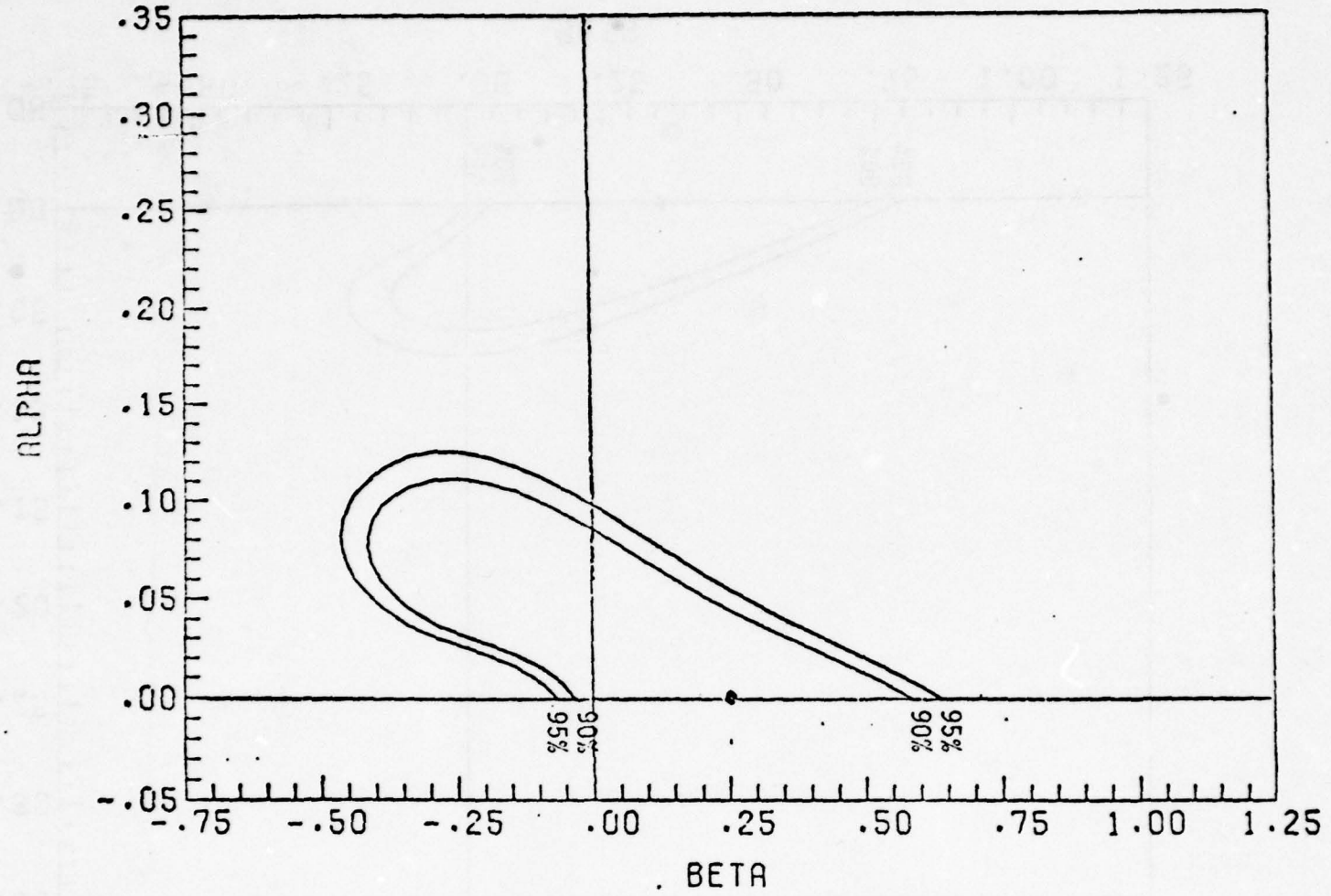


Figure 8(c) Approximate confidence regions for (β,α) (Quantity of milk in gallons for cows with percentages of butter fat between 3.00 and 3.25)

All the confidence regions in Figure 8 are reasonably close to the center of the plane, suggesting that the data sets could be random samples from distributions not too far away from the normal. It is interesting to notice that the analysis suggests that the distribution of the yields of the butter fat for cows at age 7 is a short-tailed distribution. This runs contrary to the common belief that most data sets have heavy-tailed distributions. Cox (Discussion of Stigler's paper, 1977), however, has said that according to his experience, short-tailed distributions could arise as well as heavy-tailed distributions.

These three sets of data are different in the nature from those discussed before, all three sets of data being measurements made on members of a particular population at approximately the same time, while the sets discussed previously are measurements or observations made sequentially over a long period of time. The latter type suffers from the effect of secular inhomogeneity but the former does not. This can partly explain why Tocher's first three sets of data are quite close to a normal distribution.

Due to the high cost involved in collecting the data, the Scottish Milk Record Association was not able to make daily observations of milk for each cow. The morning and evening milk of each cow was tested at intervals ranging from 14 to 28 days. The operation was repeated after an interval of about fourteen days The quantity of milk and the percentage of butter fat yielded

by each cow in the interval were assumed to be those given on the date of the previous test. So the yield of milk determined by the Scottish Milk Record Association is not the actual yield but an estimated one. During the period of 1911-13, some daily records were available from private sources. Tocher proposed to compare these actual yields with those estimated by the Association to see if the estimate is any good. The data are available on 107 cows, and we will take the difference of yields of milk (Association - Private) as observations and examine the distributional property of these 107 observations. A plot of these data is given in Figure 9. $(\hat{\theta},\hat{\sigma},\hat{\beta},\hat{\alpha})$ for these 107 data are (.10, 11.09, .20, .11), approximate confidence regions are shown in Figure 10. The confidence region is very large because the sample size is relatively small, and it is elongated,as expected.

6. Wiebe's Agricultural Data

Wiebe (1935) gives the grain yield of 1,500 rows of Federation wheat in a uniformity trial. He studied the variation and correlation in grain yield among the 1,500 wheat nursery plots to seek for criteria for better seeding arrangement in future field trials. Barbacki and Fisher (1936) used his data to examine the relative precision of systematic and randomized experimental arrangements. In field trials of this kind it is well known that results from plots close together are inevitably

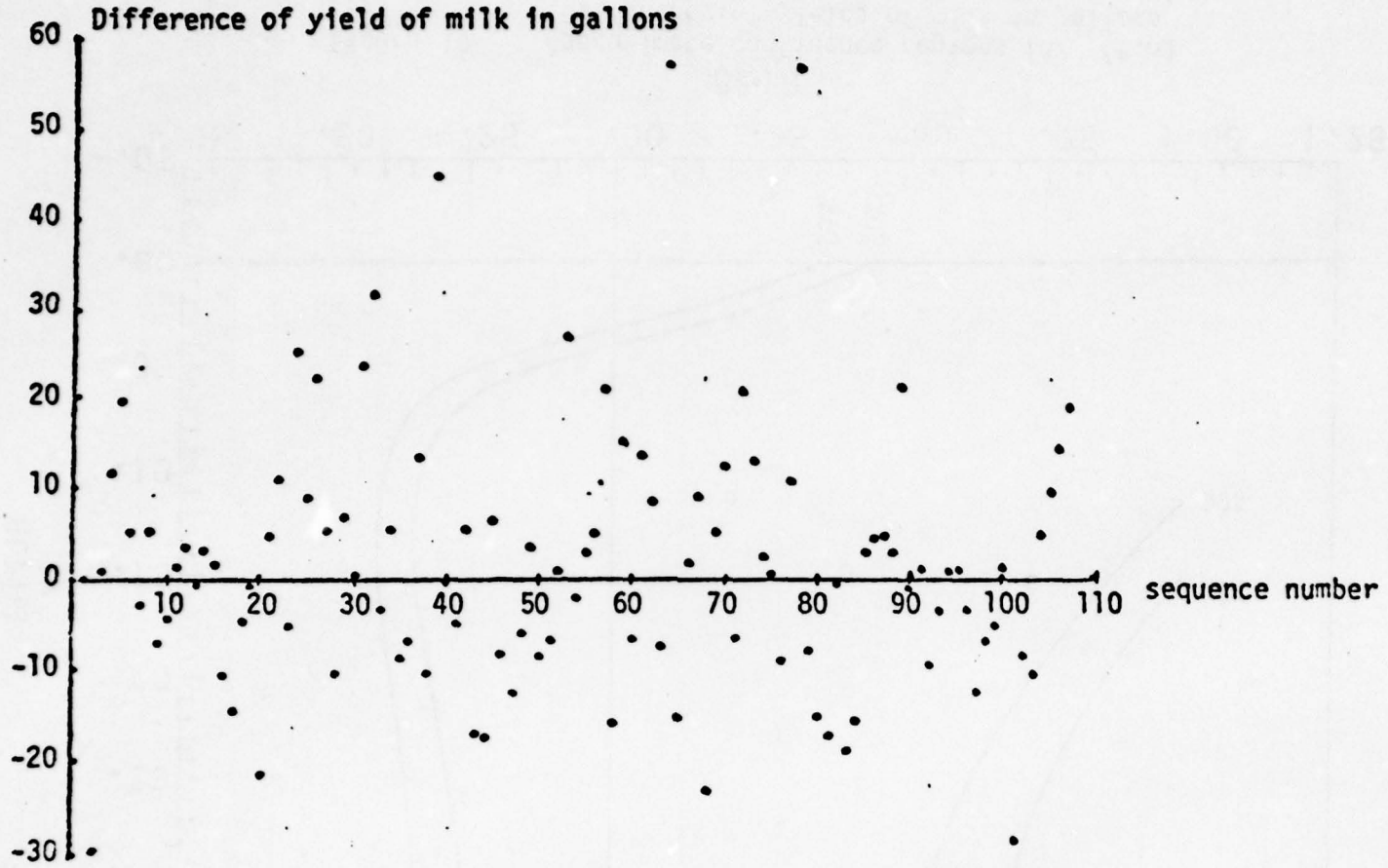


Figure 9 Difference of the yield of milk between Association and Private records.

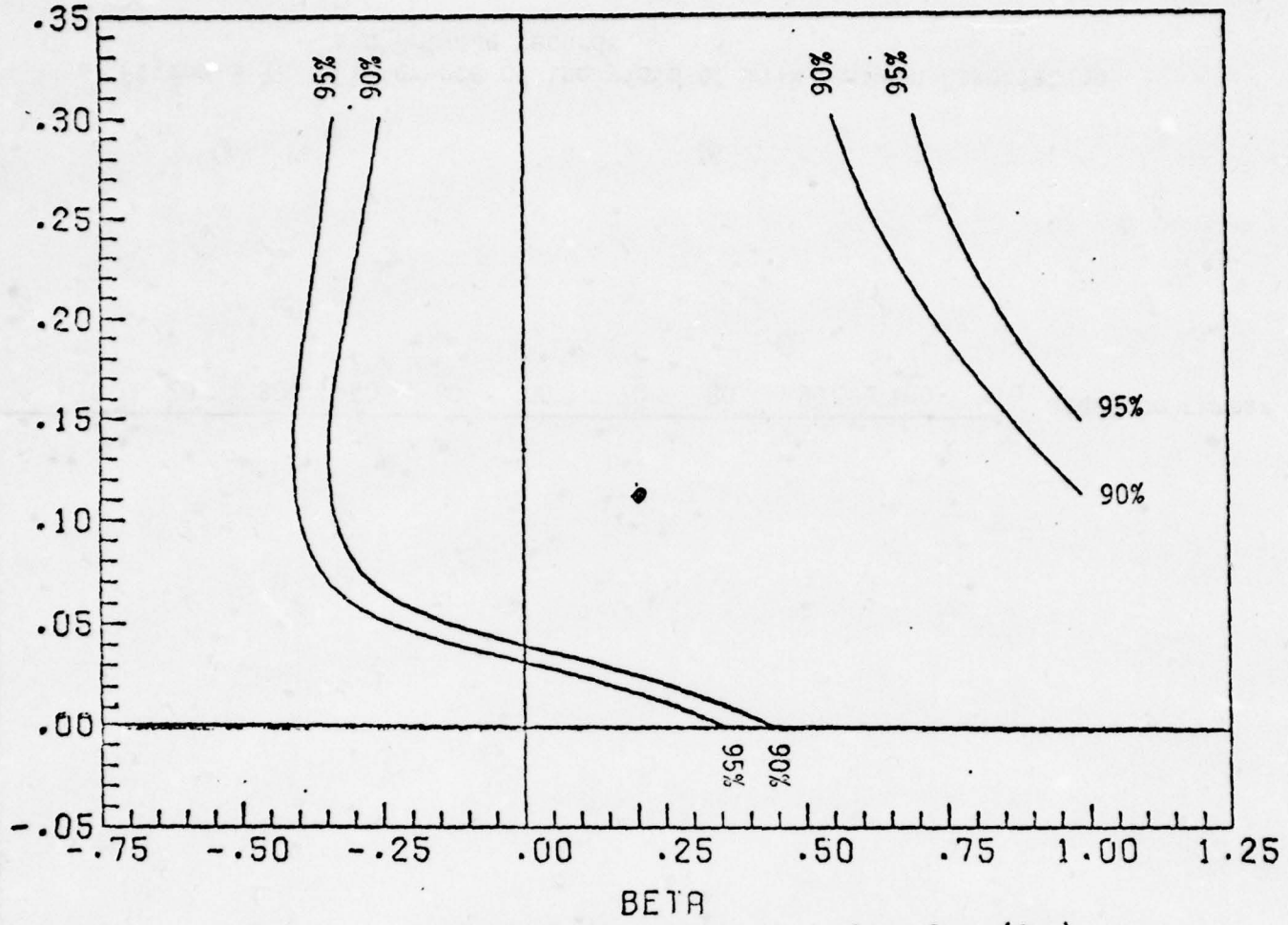


Figure 10 Approximate confidence regions for (β,α) (Difference of yields of milk in gallons between Association and private records)

correlated, and blocking and randomization were introduced to ensure proper isolation of the real effects of treatments in these circumstances.

This 1,500 rows of wheat were grown in the summer of 1927 on the Aberdeen Substation, Aberdeen, Idaho. The 1,500 rows were grown in 12 series, each having 125 rows. Figure 11 shows the general arrangement of the crop. The rows were 12 inches apart and 15 feet long. The plot was seeded on April 18 with a grain drill that sowed eight rows at a time. The individual rows were harvested in August and threshed with a small nursery thresher. The grain yields, recorded in grams per row are shown in Table . A contour map of the yield obtained in the plot given by Wiebe is shown in Figure 12 (The contour map is not obvious from the data; Wiebe must have used some kind of smoothing process which is not given in his paper). Sequence plots of all twelve series are shown in Figure 13. If we ignore the obvious serial effects and treat these series as random samples drawn from some distribution, then for any data set we can estimate θ, σ, β and α using the prodecure described on page 4. Thus for series 1, we find $(\hat{\theta},\hat{\sigma},\hat{\beta},\hat{\alpha}) = (651.6, 79.7, \underline{0.45}, \underline{0.0})$ with the approximate confidence regions shown in Figure 14. Although the maximum likelihood estimates suggest an approximate uncontaminated somewhat heavy-tailed parent distribution, the confidence regions for (β,α) still include the normal distribution.

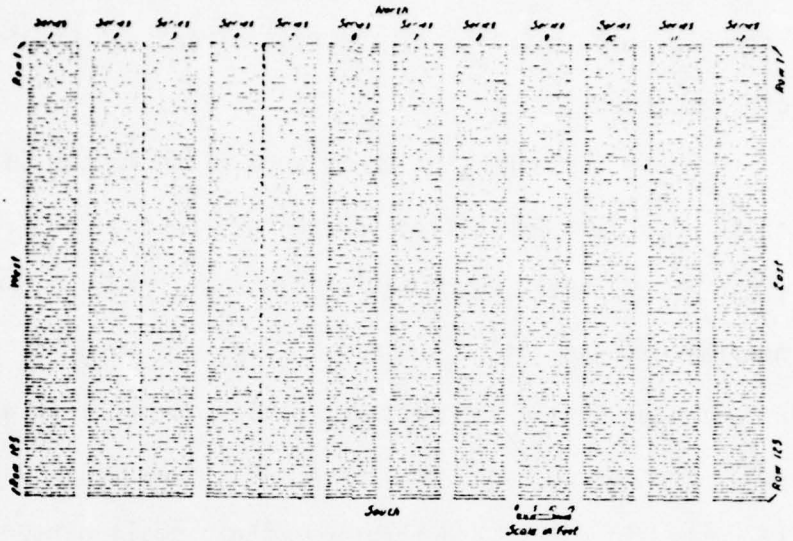


Figure 11 Field plan of uniformity experiments showing general arrangement of rows. Series 2 and 3, 4 and 5, were grown as 30 foot rows but were cut in 15 foot lengths at harvest time.

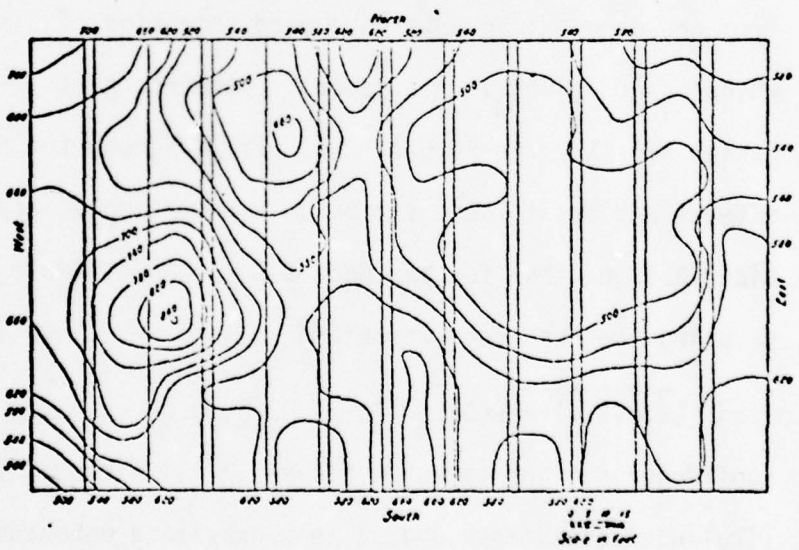


Figure 12 Wiebe's contour map of the grain yields obtained. The lines are at 40-g intervals.

Table 2 Yield of grain from each of 1,500 15-foot rows of wheat, 12 inches apart, grouped in 12 series of 125 rows each, Wiebe's data.

Series 1	Series 2	Series 3	Series 4	Series 5	Series 6	Series 7	Series 8	Series 9	Series 10	Series 11	Series 12
715	595	560	560	615	610	540	515	517	665	560	612
770	710	655	675	710	690	565	565	590	574	511	618
760	715	690	690	655	725	695	640	645	705	644	705
685	615	685	555	595	630	550	525	503	614	573	570
755	720	670	565	545	620	545	525	495	565	509	612
745	670	565	560	520	710	560	545	558	567	600	664
645	690	520	520	450	677	555	505	520	535	611	578
565	495	455	470	445	555	550	459	420	461	531	559
560	540	450	540	505	595	545	530	454	535	453	600
685	720	610	540	555	645	625	525	534	563	616	638
755	810	665	570	525	715	650	560	613	617	742	657
640	635	565	495	470	635	550	515	509	493	635	567
725	655	530	455	465	560	550	450	455	503	519	555
715	775	615	545	530	645	565	510	507	501	561	537
700	705	555	440	455	630	555	425	475	445	512	552
640	635	495	435	465	545	475	435	499	516	455	559
629	635	495	445	455	629	505	455	419	591	545	537
750	630	555	455	510	575	550	470	427	545	562	582
760	675	610	540	530	675	580	460	513	559	595	675
630	645	490	445	440	565	490	429	469	542	474	563
625	540	495	415	445	509	440	450	455	465	461	533
670	670	560	465	445	590	555	450	516	570	503	623
655	630	540	455	450	560	465	435	466	454	455	503
570	535	525	455	415	510	460	429	368	452	425	445
575	495	540	480	395	555	440	450	437	541	525	514
690	645	595	515	470	590	515	445	528	588	517	533
730	730	710	565	495	640	565	550	494	591	594	597
620	635	595	495	435	530	490	470	405	521	454	517
620	600	530	455	425	545	505	455	442	494	505	578
665	665	625	520	450	525	550	465	467	595	563	634
620	635	530	440	450	590	470	445	510	531	478	514
615	525	610	515	395	540	455	520	495	554	445	533
595	625	600	440	435	510	515	525	509	522	495	514
640	560	630	495	490	610	525	515	574	525	495	518
625	650	695	560	515	665	570	525	613	560	532	525
630	610	545	465	450	555	500	525	495	477	450	507
595	555	540	505	465	540	425	490	457	479	444	507
645	635	600	560	525	645	515	520	477	510	495	533
705	660	595	495	495	595	490	505	497	499	509	533
575	530	600	490	410	500	490	505	459	491	438	578
530	555	570	490	425	490	490	495	425	443	449	499
715	660	665	505	535	640	495	490	474	491	495	533
795	715	740	540	565	565	535	560	518	513	541	545
650	635	620	495	510	585	520	475	444	485	509	469
615	580	505	490	505	555	490	470	455	442	474	484
690	645	640	600	565	645	550	550	494	474	592	548
675	615	640	505	550	615	535	590	465	456	513	563
770	625	665	570	575	625	540	475	534	456	467	545
695	700	770	540	515	625	530	495	508	454	422	600
650	660	705	570	545	645	570	540	494	540	532	587
720	730	825	665	605	665	520	550	564	546	537	630
670	640	710	540	535	575	495	455	512	447	465	600
655	590	675	575	515	570	490	470	490	457	472	569
780	790	855	625	575	655	530	540	575	502	503	592
750	705	770	575	560	565	510	435	446	470	502	570
665	630	695	620	550	520	435	490	429	465	534	619
665	605	660	570	560	560	445	455	451	453	523	627
735	760	790	645	605	595	540	455	466	447	476	645
795	795	850	700	665	635	540	470	521	542	564	759
645	665	720	610	520	550	490	440	425	465	462	598
630	695	665	630	510	575	445	430	483	483	450	627
695	840	875	640	575	625	570	495	514	540	496	717
695	790	790	575	520	630	545	470	455	496	494	619
705	745	840	710	640	610	500	540	494	478	434	574
655	735	840	640	610	565	560	450	489	542	450	570
640	850	795	660	560	615	565	460	451	492	504	630
790	765	840	770	660	640	575	490	465	495	529	552
670	690	785	645	510	560	535	425	474	455	454	555
695	790	770	665	620	545	555	540	509	451	439	615
660	825	950	730	660	620	629	510	501	440	524	694
705	695	850	635	540	660	535	435	429	457	451	649
610	720	705	615	540	625	595	540	379	442	502	638
640	735	805	665	535	605	560	430	427	441	494	615
690	855	905	770	615	650	615	495	501	444	453	645
715	765	945	820	695	750	655	530	511	560	510	762

Table 2 continued

Series 1	Series 2	Series 3	Series 4	Series 5	Series 6	Series 7	Series 8	Series 9	Series 10	Series 11	Series 12
685	715	825	715	550	650	655	505	414	451	542	642
695	690	790	695	595	625	530	525	425	421	475	514
725	645	765	820	655	675	640	545	474	513	589	597
645	670	840	740	615	655	555	495	433	455	591	567
725	825	820	640	690	650	525	525	451	454	510	567
650	865	835	690	645	640	590	520	476	507	525	563
695	860	790	725	615	710	650	525	495	459	546	603
735	810	805	765	620	700	770	635	505	525	514	615
645	745	815	700	605	615	605	550	455	474	583	567
630	810	790	635	535	650	735	535	495	454	535	615
640	745	840	730	645	650	775	550	550	561	565	619
620	730	775	660	610	610	660	515	537	535	553	570
620	745	660	565	520	635	610	450	476	550	562	575
560	675	660	605	525	605	615	515	454	518	558	550
625	700	720	645	645	560	615	540	516	618	615	604
700	765	725	615	640	700	710	550	614	627	620	645
685	765	695	690	570	615	660	575	555	602	580	590
625	650	640	650	605	585	650	565	532	624	571	595
745	790	675	690	605	765	725	605	636	666	568	632
660	670	630	660	645	650	605	610	622	572	587	625
655	730	615	650	640	645	655	620	611	642	605	630
625	700	675	720	605	660	720	615	595	555	645	578
670	735	645	620	705	695	660	567	603	541	550	630
670	823	695	665	735	715	740	650	658	624	725	675
555	670	595	740	580	685	640	605	553	550	635	634
650	665	595	625	595	590	560	575	592	559	564	518
625	760	625	645	605	635	660	625	672	636	577	645
623	740	575	565	595	610	650	595	515	641	560	623
580	650	570	575	470	445	670	645	567	634	546	524
615	705	550	515	550	465	610	570	565	652	515	623
675	690	615	610	550	625	745	560	564	616	535	668
720	820	720	650	560	615	770	645	615	755	625	713
640	760	640	690	645	540	695	550	615	679	560	570
630	690	715	710	605	575	695	625	564	665	604	592
625	830	795	755	610	630	775	600	614	709	730	619
670	665	690	735	560	560	670	610	556	592	458	634
645	555	605	665	540	615	610	525	540	650	594	734
670	570	555	655	475	625	725	525	565	540	623	745
650	695	650	690	585	660	760	655	658	621	576	615
555	650	715	690	605	665	775	655	652	720	642	728
620	530	690	605	650	545	585	630	551	670	622	623
650	675	595	660	675	545	605	550	524	554	605	552
600	610	560	660	555	625	620	665	551	613	562	565
625	560	570	570	655	605	605	530	544	645	612	537
555	550	540	630	650	565	545	515	467	636	563	510
560	590	560	575	580	675	610	555	582	541	572	582
640	610	605	620	650	585	650	550	577	612	654	630
670	555	635	765	650	675	765	620	596	705	657	650
555	590	560	655	670	555	550	555	537	585	549	619
655	630	510	640	650	600	670	615	653	594	616	784

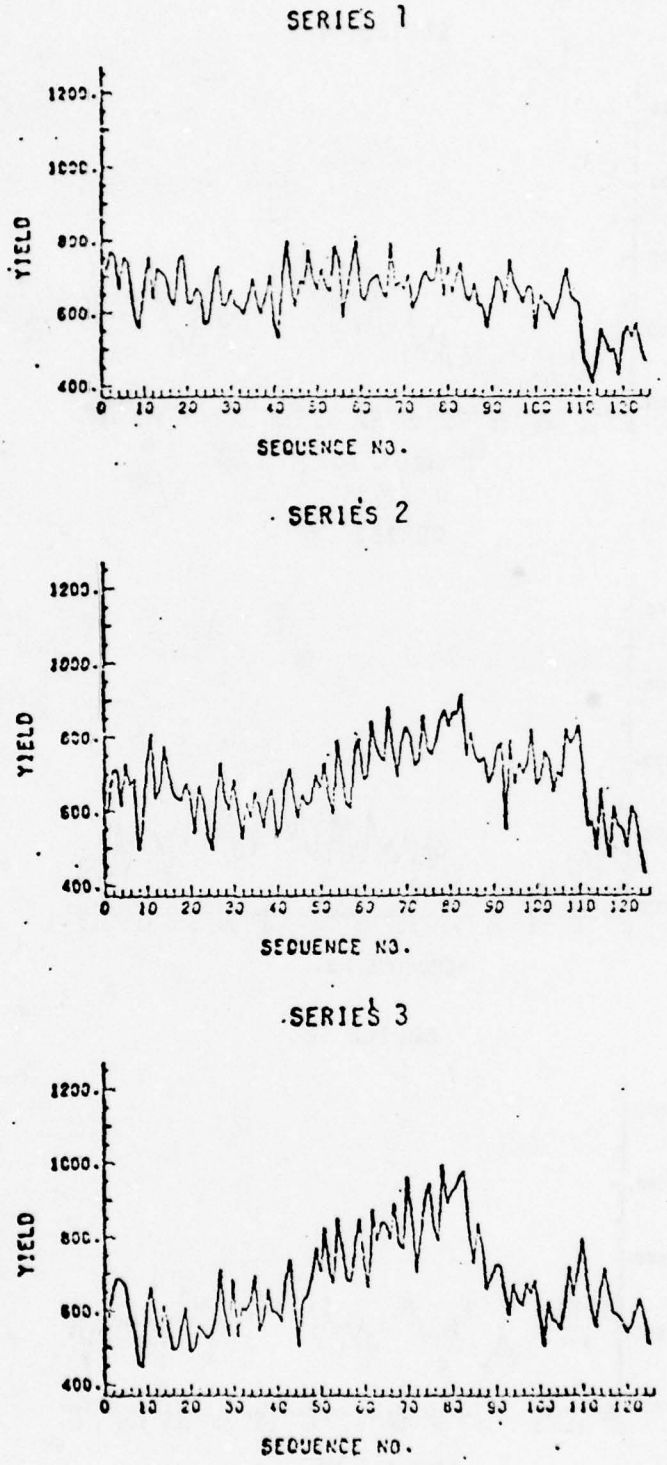
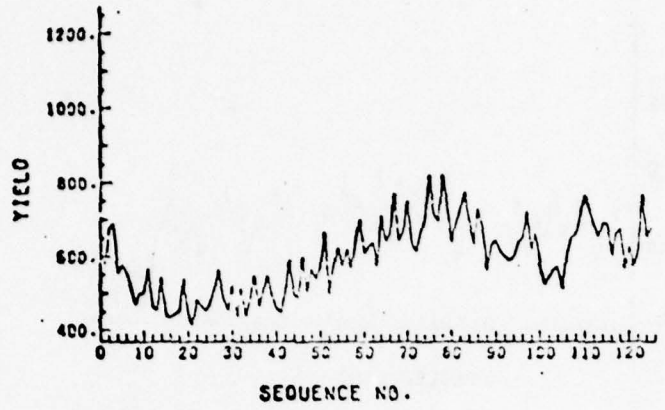
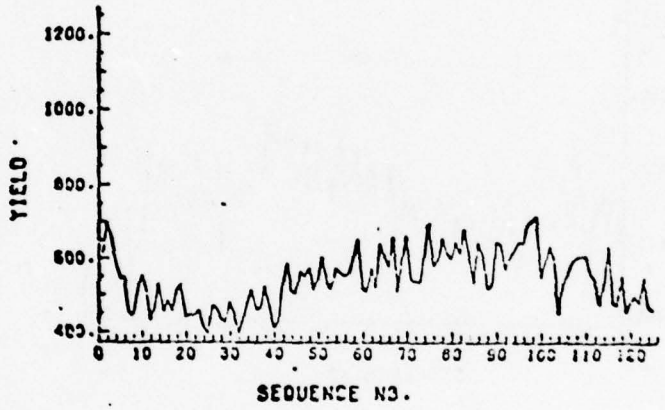


Figure 13 Sequence plot of Wiebe's 12 series given in Table 3.2.

SERIES 4



SERIES 5



SERIES 6

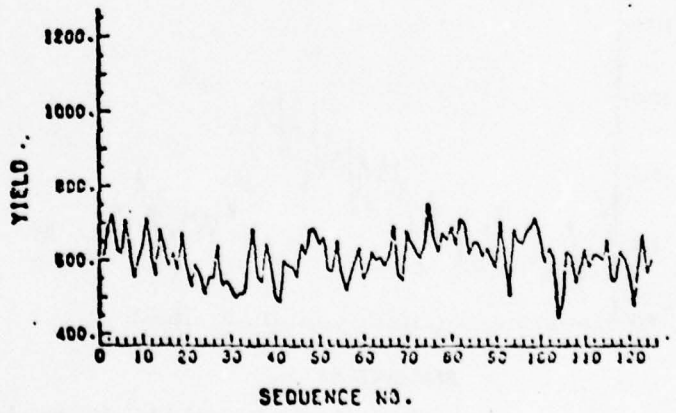
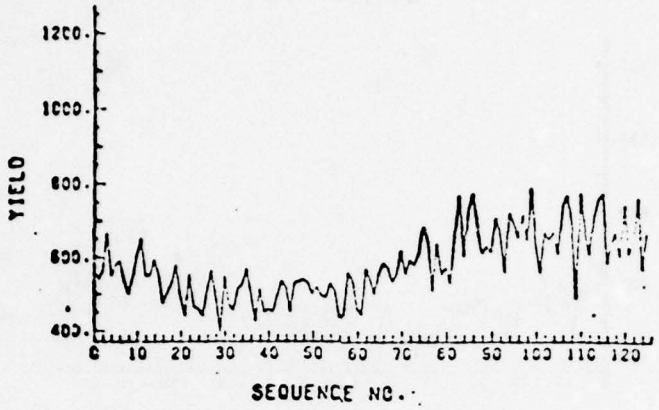
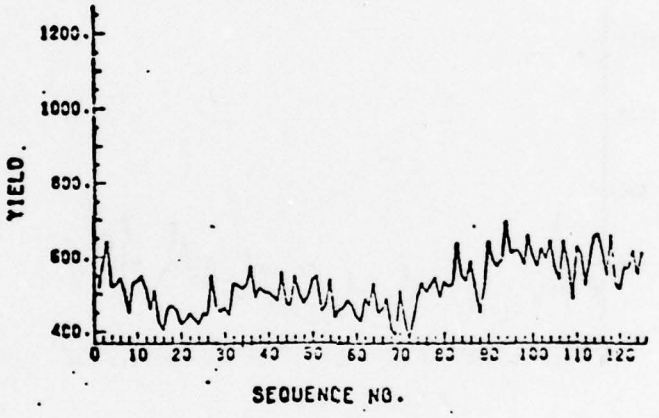


Figure 13 continued

SERIES 7



SERIES 8



SERIES 9

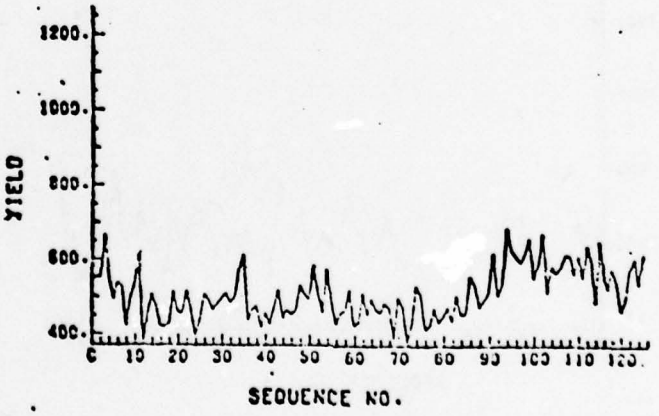
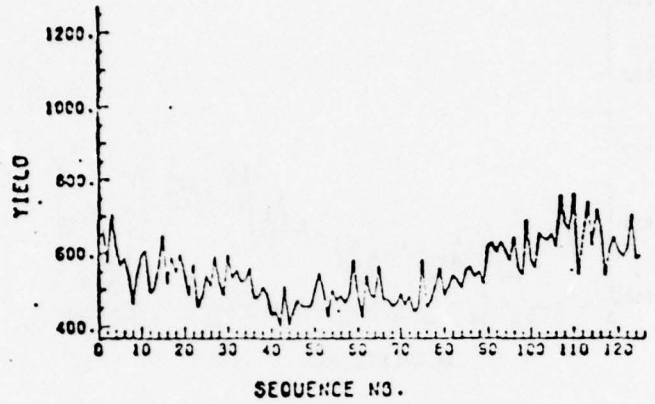
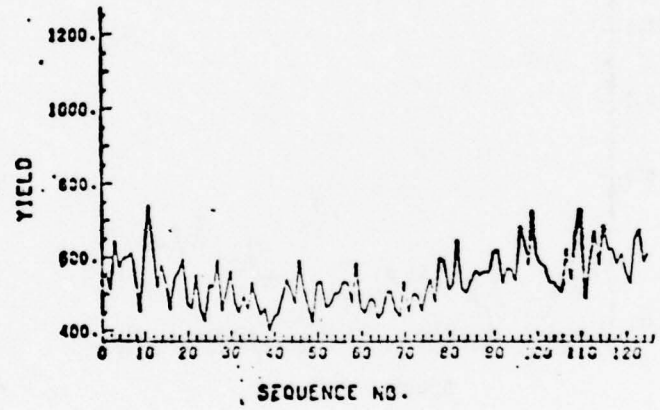


Figure 13 continued

SERIES 10



SERIES 11



SERIES 12

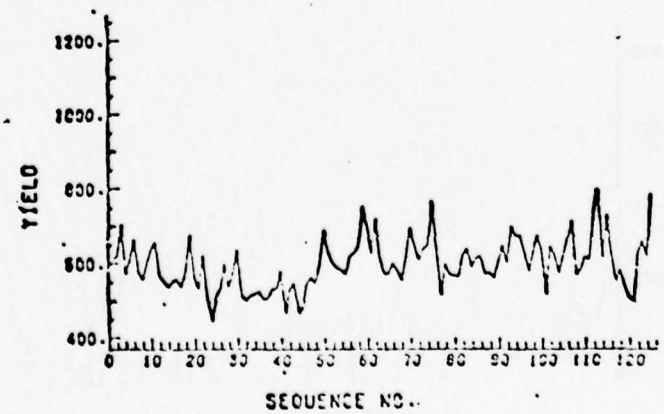


Figure 13 continued

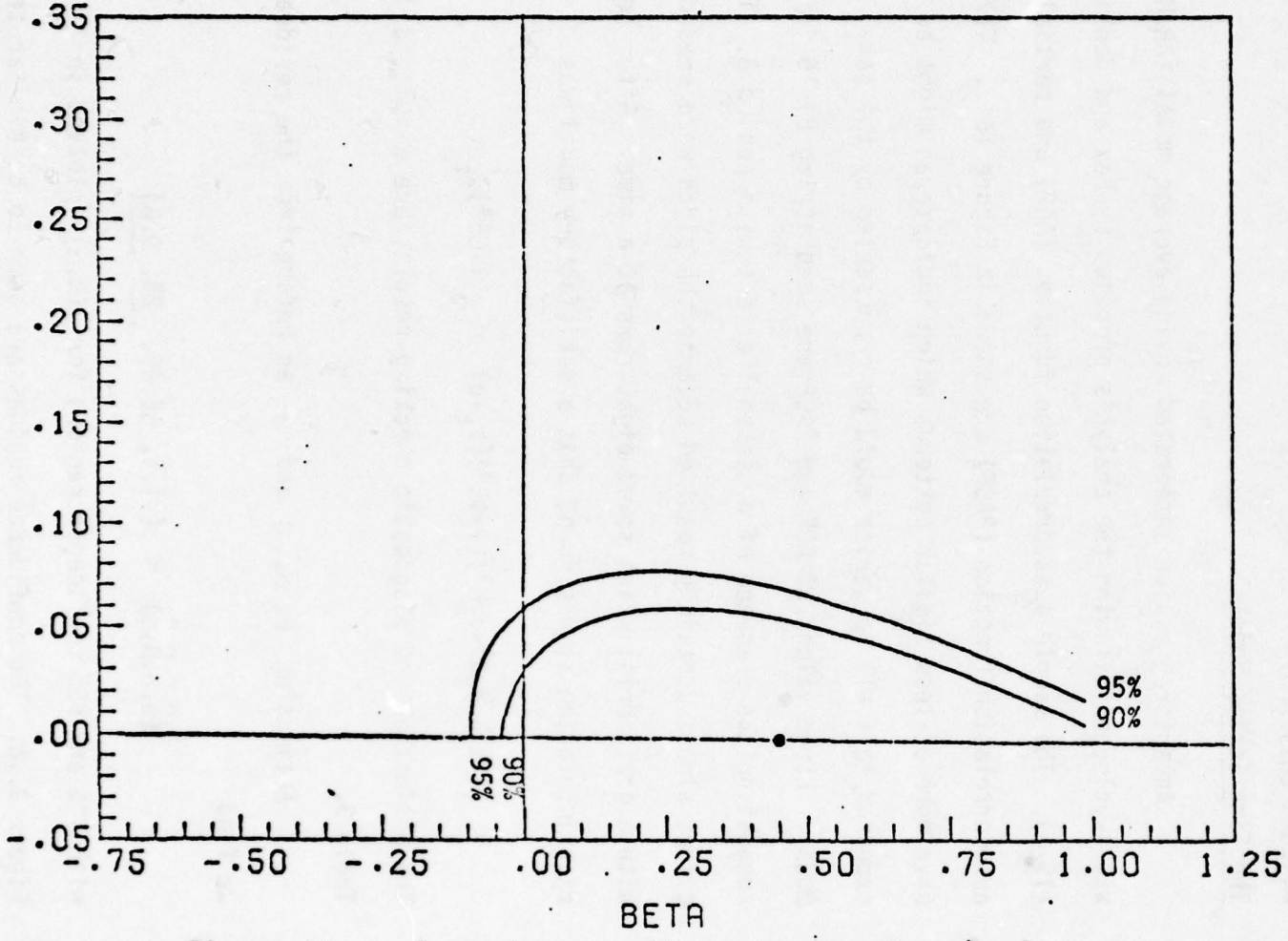


Figure 14 Approximate confidence regions for (β,α) (Wiebe's agricultural data, series 1)

More complicated models which take into account the serial correlations were also fitted to this set of data. We discuss this now.

The Time Series Model

An autoregressive integrated moving average model (ARIMA) was developed following the analysis proposed by Box and Jenkins (1976). The sample autocorrelation function (ACF) and partial autocorrelation function (PACF) are shown in Figure 15 . They have more or less similar patterns which indicate, as might be expected, that all 12 series could be represented by the same model. Large values of ACF and PACF are seen at lag 8, 16 suggesting the presence of a seasonal effect with period 8. This effect almost certainly resulted because the plots were seeded with a grain drill which sowed eight rows at a time. After some experimentation it was found that a satisfactory model was

$$(1-\phi_1 B-\phi_2 B^2-\phi_3 B^3)(1-\Phi B^8)(z_t-u) = (1-\Theta B^8)a_t .$$

The estimation and diagnostic checking results are displayed in Table 3.

Estimating θ, σ, β and α as before from the residuals, we find

$$(\hat{\theta},\hat{\sigma},\hat{\beta},\hat{\alpha}) = (-1.1,\ 56.45,\ \underline{.25},\ \underline{0.0})$$

with 90% and 95% confidence regions for (β,α) plotted in Figure 3.16. The confidence regions are seen to be similar to those found for z_t.

Series 1

Series 2

Series 3

Figure 15(a) ACF for Wiebe's 12 series given in Table 2.

Series 4

Series 5

Series 6

Figure 15(a) continued

Series 7

Series 8

Series 9

Figure 15(a) continued

Series 10

Series 11

Series 12

Figure 15(a) continued

Series 1

Series 2

Series 3

Figure 15(b) PACF for Wiebe's 12 series given in Table 2.

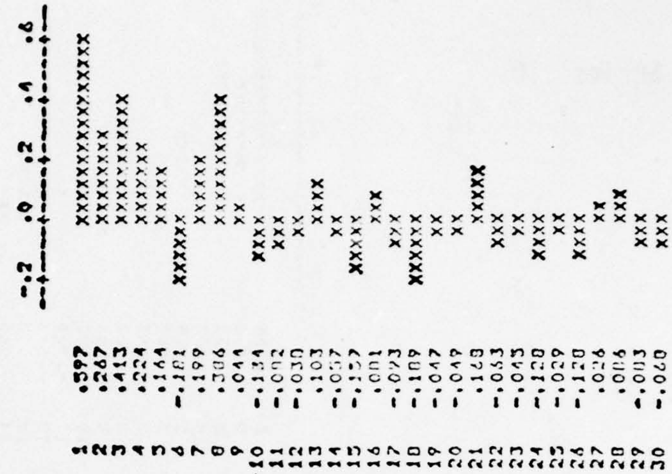
Series 4

Series 5

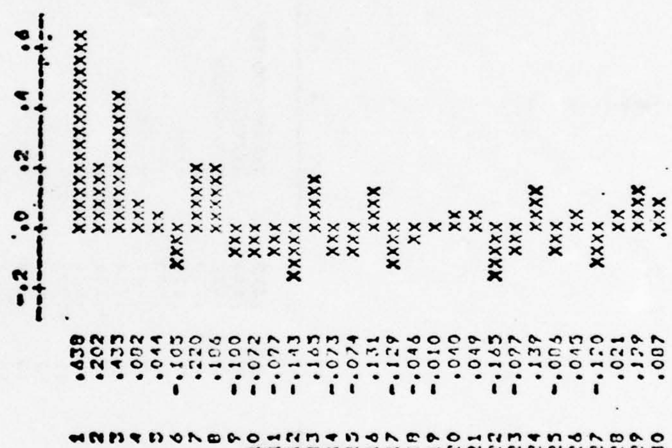
Series 6

Figure 15(b) continued

Series 7



Series 8



Series 9

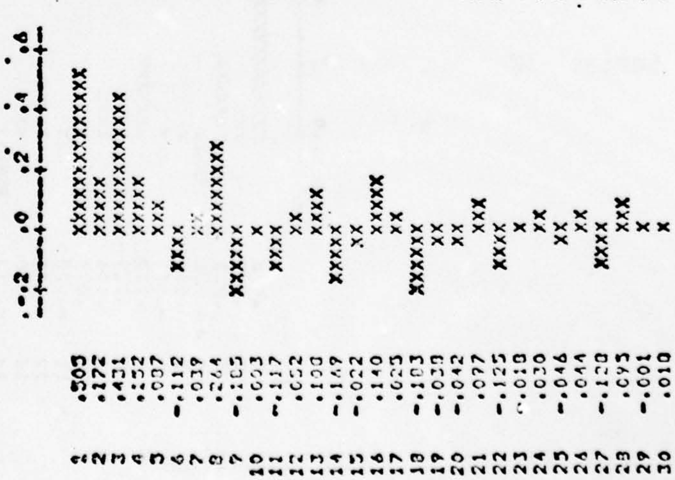
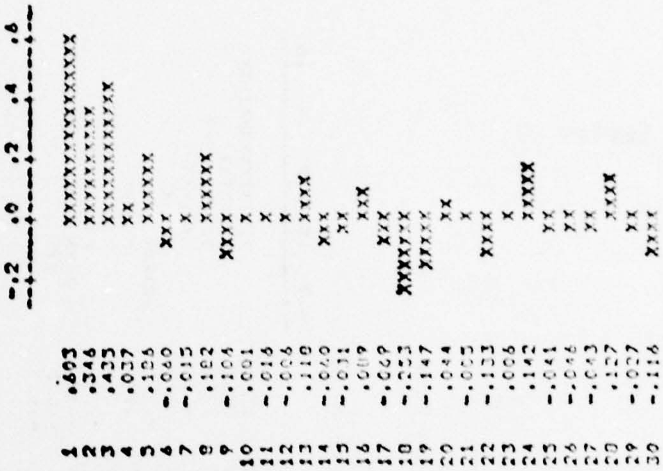
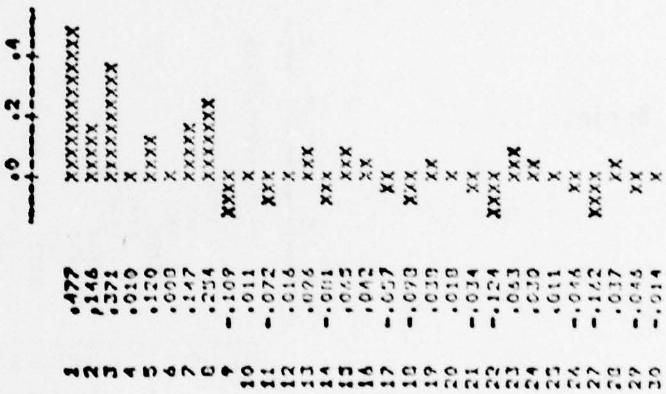


Figure 15(b) continued

Series 10



Series 11



Series 12

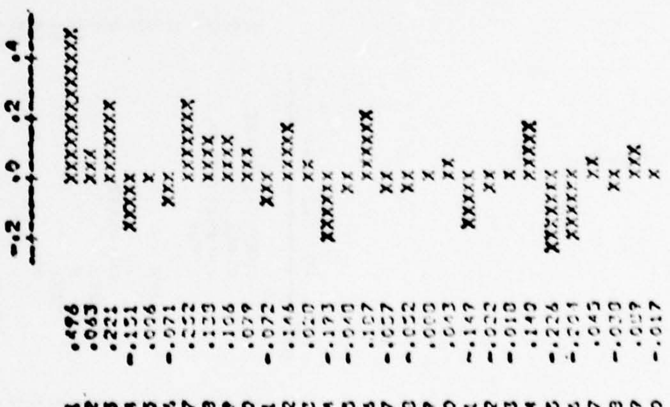


Figure 15(b) continued

Table 3. Estimation results for model $(1-\phi_1 B-\phi_2 B^2-\phi_3 B^3)(1-\Phi_8 B^8)(z_t-u) = (1-\Theta_8 B^8)a_t$ for Wiebe's twelve series.

Series	estimated parameters						M.S.E.($\times 10^{-2}$) 108 d.f.	Diagnostic checking: $\chi^2_{.05}(14) = 23.69$, $\chi^2_{.10}(14) = 21.06$	ACF, PACF	lack of fit
	ϕ_1	ϕ_2	ϕ_3	Φ_8	u	Θ_8				
1	.55	.04	.17	.89	577.63	.73	26.89	21.67	not very good	boundary
2	.36	.27	.20	.79	630.28	.54	46.06	11.72	fair	
3	.68	.04	.14	.87	687.81	.48	34.09	26.88		yes
4	.37	.22	.20	.95	946.15	.82	20.84	26.01		yes
5	.52	.01	.25	.90	610.60	.65	19.68	13.15	random	
6	.51	-.03	.29	.95	595.65	.87	18.15	13.91	fair	
7	.22	.27	.28	.98	1060.70	.76	29.04	10.79	fair	
8	.41	-.00	.39	.95	726.38	.76	16.38	13.24	random	
9	.34	.04	.44	.87	571.52	.70	19.73	8.43	random	
10	.29	.13	.41	.87	615.28	.68	22.41	17.87	fair	
11	.31	.00	.37	.47	534.67	.19	27.61	11.52	fair	
12	.55	.00	.17	.93	643.77	.90	27.41	45.34		yes

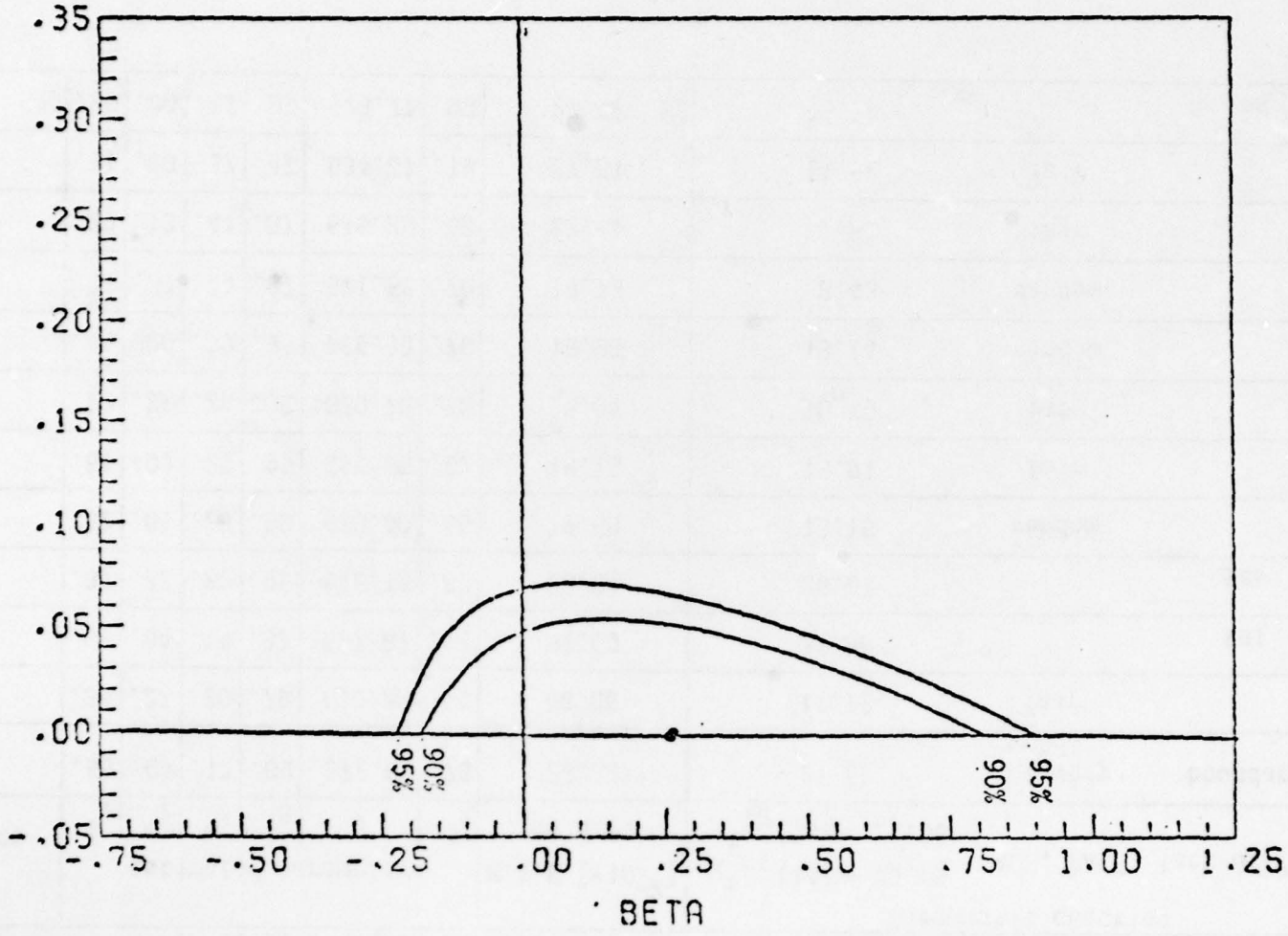


Figure 16. Approximate confidence regions for (β,α)
(Wiebe's data: residuals from time series model)

In general, the relationship between the kurtosis for z_t and a_t is readily established for a stationary process $\phi(B)z_t = \theta(B)a_t$.

Consider first a moving average process

$$z_t = (1-\theta_1 B-\theta_2 B^2-\ldots-\theta_q B^q)a_t \qquad \theta_i\text{'s not all zero}$$

$$= a_t-\theta_1 a_{t-1}-\theta_2 a_{t-2}-\ldots-\theta_q a_{t-q} .$$

If $a_t, a_{t-1}, \ldots, a_{t-q}$ are independently and identically distributed with cumulants $k_i(a_t)$ and coefficient of kurtosis $\lambda_4(a_t)$, then the i-th cumulant for z_t will be

$$k_i(a_t)+\theta_1^i k_i(a_t)+\theta_2^i k_i(a_t)+\ldots+\theta_q^i k_i(a_t) = (1+\theta_1^i+\ldots+\theta_q^i)k_i(a_t).$$

The coefficient of kurtosis $\lambda_4(z_t)$ is then

$$\lambda_4(z_t) = \frac{k_4(z_t)}{k_2^2(z_t)} = \frac{(1+\theta_1^4+\ldots+\theta_q^4)k_4(a_t)}{(1+\theta_1^2+\ldots+\theta_q^2)^2 k_2^2(a_t)} = \frac{1+\theta_1^4+\ldots+\theta_q^4}{(1+\theta_1^2+\ldots+\theta_q^2)^2}\lambda_4(a_t) .$$

It is easy to see since $\dfrac{1+\theta_1^4+\ldots+\theta_q^4}{(1+\theta_1^2+\ldots+\theta_q^2)^2} < 1$, that z_t is more normal than a_t and in particular, z_t is normally distributed provided that a_t is so distributed.

Any stationary model $\phi(B)z_t = \theta(B)a_t$ can be written in the form

$$z_t = \sum_{j=0}^{\infty} \psi_j a_{t-j} \quad \text{where} \quad \sum_{j=0}^{\infty} \psi_j \quad \text{converges.}$$

It then follows that in general

$$\lambda_4(z_t) = \frac{\left(\sum_{i=0}^{\infty} \psi_i^4\right) k_4(a_t)}{\left(\sum_{i=0}^{\infty} \psi_i^2\right)^2 k_2^2(a_t)} = \frac{\left(\sum_{i=0}^{\infty} \psi_i^4\right)}{\left(\sum_{i=0}^{\infty} \psi_i^2\right)^2} \lambda_4(a_t)$$

and $\dfrac{\left(\sum_{i=0}^{\infty} \psi_i^4\right)}{\left(\sum_{i=0}^{\infty} \psi_i^2\right)^2} < 1$. The previous conclusions for the moving average, therefore, hold for any stationary series.

What we have is evidently a central limit effect which makes z_t more normal than a_t. In the important special case of a first order autoregressive process

$$(1-\phi B)z_t = a_t \qquad |\phi| < 1$$

$$z_t = (1+\phi B+\phi^2 B^2+\phi^3 B^3+\ldots)a_t$$

$$= a_t+\phi a_{t-1}+\phi^2 a_{t-2}+\phi^3 a_{t-3}+\ldots$$

and $\lambda_4(z_t) = \dfrac{\sum_{i=0}^{\infty} (\phi_i)^4}{\left(\sum_{i=0}^{\infty} (\phi_i)^2\right)^2}$ $\lambda_4(a_t) = \dfrac{\frac{1}{1-\phi^4}}{\left(\frac{1}{1-\phi^2}\right)^2}$ $\lambda_4(a_t) = \dfrac{1-\phi^2}{1+\phi^2}\lambda_4(a_t)$.

The larger ϕ is, the smaller $\dfrac{1-\phi^2}{1+\phi^2}$ will be and the stronger will be the central limit effect.

The situation is different for a nonstationary process. This may be demonstrated by a simulation study using a random walk model.

(i) Generate n random normal deviates $a_1,\ldots,a_n$.

(ii) Calculate $z_t = \sum_{i=1}^{t} a_i$.

(iii) Calculate $\dfrac{n\sum_{i=1}^{n}(z_i-\bar{z})^4}{\left(\sum_{i=1}^{n}(z_i-\bar{z})^2\right)^2} - 3$ which is an estimate of the coefficient of kurtosis λ_4.

With $n = 100$, five such simulations gave estimates of λ_4 of -.44, -.61, -.56, -.28, -.61. With $n = 200$, three simulations gave estimates of λ_4 of -1.14, -.95 and -2.0.

For Wiebe's data, nonstationary models were tried but did not fit as well as the stationary model. The observed result for this data that the tail behavior for z_t and a_t is similar will be expected when a_t is nearly normally distributed and a stationary model is appropriate.

An Alternative Approach: the Block Model

Traditionally, the kind of design that has been used for agricultural data employs small blocks. Since the drill machine

sowed eight rows at a time, every eight rows can conveniently be viewed as a block and associated with a block effect B_j. A treatment effect D_i could be associated with the position of the drill.

A plausible model is then

$$z_t = \mu + D_{i_t} + B_{j_t} + \varepsilon_t \qquad t = 1,\ldots,125,\ 1 \le i_t \le 8,\ 1 \le j_t \le 16$$

where $i_t = t_{\text{mod } 8} = t - [(t-1)/8]^* \times 8$

$j_t = [(t-1)/8] + 1$

and $\Sigma D_{i_t} = 0 \qquad \Sigma B_{j_t} = 0.$

The "design" thus produced is systematic and not randomized. However, we might allow for serial correlation in the errors by letting ε_t be an autoregressive process

$$(1-\phi B)\varepsilon_t = a_t \quad \text{where } a_t\text{'s are i.i.d. } N(0,\sigma^2).$$

The model now becomes

$$z_t = \mu + D_{i_t} + B_{j_t} + \varepsilon_t \quad \text{with } (1-\phi B)\varepsilon_t = a_t,\ a_t \text{ i.i.d. } N(0,\sigma^2). \qquad \text{Model (1)}$$

The parameters can now be estimated using nonlinear least squares.

Alternatively we can apply linear least squares estimation in the model

---

* $[x]$ means the integer part of x.

$$Y_t = (1-\phi)\mu + \nabla_{i_t} + \beta_{j_t} + a_t \quad \text{where}$$

$$Y_t = z_t - \phi z_{t-1}, \quad \nabla_{i_t} = D_{i_t} - \phi D_{i_{t-1}}, \quad \beta_{j_t} = B_{j_t} - \phi B_{j_{t-1}}$$

for various ϕ values, and plot the residual mean square against ϕ. The estimate of $\hat{\phi}$ is obtained at the minimum of the curve and all the other parameters are estimated by linear least squares with this $\hat{\phi}$ value.

Figure 17 shows this estimation procedure for Series 1 which yields $\hat{\phi} = 0.26$. The estimates of all the other parameters are given in Table 4 . The ACF and PACF for the residuals are given in Figure 18 . No significant pattern is noticeable.

If we now look back to the time series models we have fitted, we see that even the best fitted model for Series 1 has a residual mean square of over 2600 which is considerably larger than the residual mean square for the present model (2078). Even if we did not allow for serial correlation for the errors and made the assumption that the errors were i.i.d. $N(0,\sigma^2)$, i.e., $\phi = 0$, the residual mean square 2216 is still relatively small although ACF and PACF for the residuals (Figure 19) do now show some pattern. It is clear, however, that the block model does very well in explaining the data.

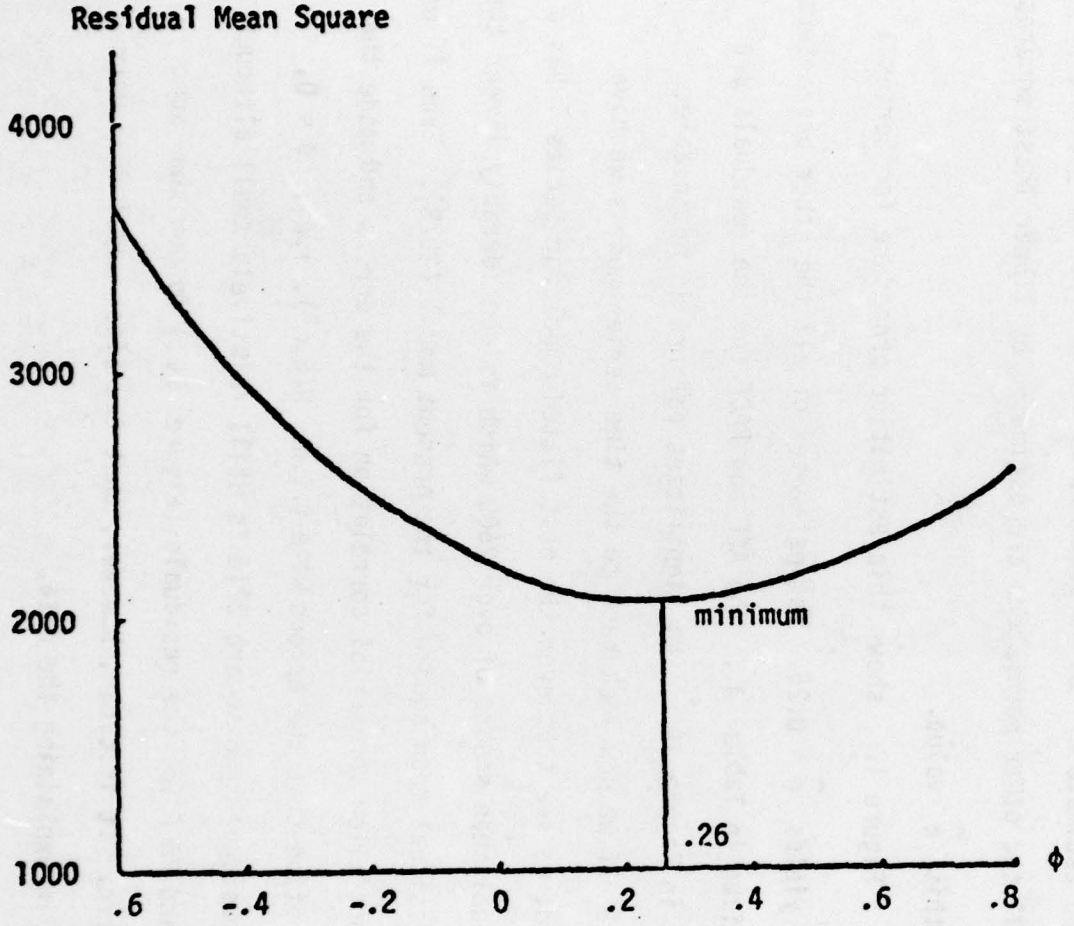
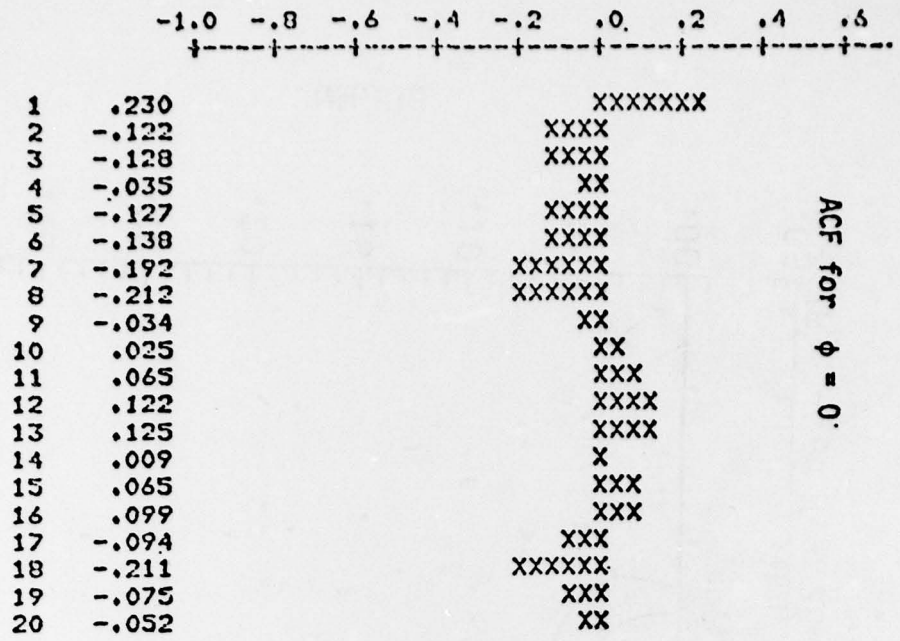


Figure 17 Residual mean square curve with the block model for Series 1.

Table 4 Estimates of the parameters in model (1)

parameter	estimated value	parameter	estimated value
ϕ	.26	B_4	.3
μ	643.63	B_5	- 4.4
D_1	- 43.9	B_6	37.0
D_2	20.9	B_7	37.6
D_3	72.7	B_8	55.9
D_4	- 14.6	B_9	31.3
D_5	- 13.6	B_{10}	54.0
D_6	27.7	B_{11}	16.8
D_7	- 11.4	B_{12}	12.2
D_8	- 37.8	B_{13}	-22.2
B_1	50.5	B_{14}	-41.7
B_2	33.5	B_{15}	-151.5
B_3	15.1	B_{16}	-124.4

Figure 18 ACF and PACF for residuals obtained from fitting Series 1 to model (1).



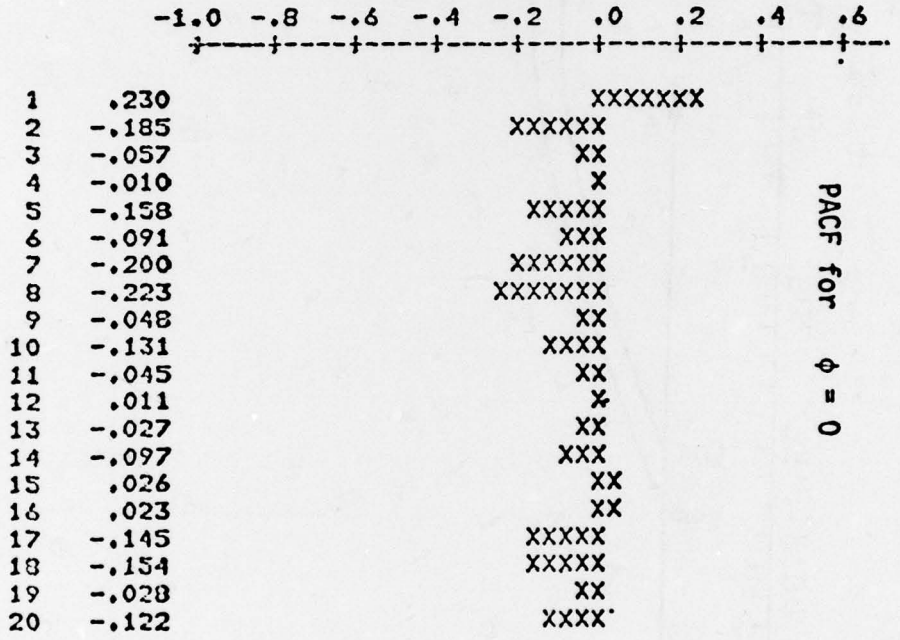


Figure 19 ACF and PACF for residuals obtained from fitting Series 1 to model (1).

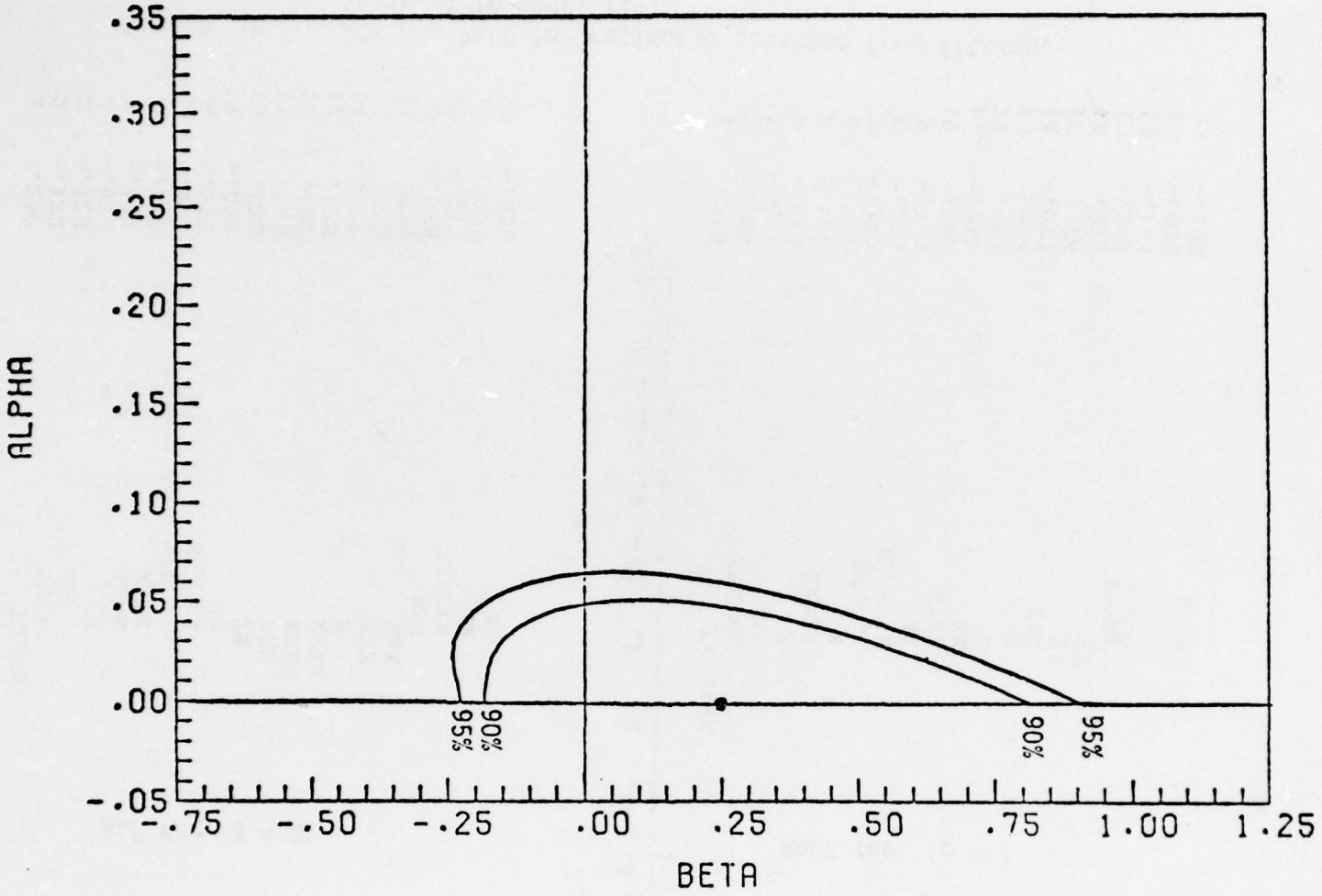


Figure 20 Approximate 90% and 95% confidence regions for (β,α) (Data: Residuals obtained from fitting Series 1 to model (1) with $\phi = 0.26$)

References

1. Bachelier, L. (1900), "Theorie de la speculation," Ann. Sci. Ec. norm. sup., Paris, Series 3, 17, 21.

2. Barbacki, S. and R. A. Fisher (1936), "A test of the supposed precision of systematic arrangements," Annals of Eugenics, 7, p 189-193.

3. Box, G. E. P. and G. M. Jenkins (1976), Time Series Analysis: Forecasting and Control, Holden-Day.

4. Lund, D. R. (1967), Parameter Estimation in a Class of Power Distributions, Ph.D. thesis, the University Wisconsin-Madison.

5. Michelson, A. A., Pease, F. G. and F. Pearson (1935), "Measurement of the velocity of light in a partial vacuum," Astrophysical Journal, 82, p 26-61.

6. Newcomb, S. (1891), "Measures of the velocity of light made under the direction of the Secretary of the Navy during the years 1880-1882," Astronomical Papers, 2, p 107-203, U. S. Nautical Almanac Office.

7. Stigler, S. M. (1977), "Do robust estimators work with real data?" The Annals of Statistics, 5, p 1055-1098.

8. Tocher, J. F. (1928), "An investigation of the milk yield of dairy cows," Biometrika, 20B, Part II, p 105-244.

9. Wichern, D. W., Miller, R. B. and D. Hsu (1976), "Changes of variance in first-order autoregressive time series models - with an application," Applied Statistics, 25, p 248-256.

10. Wiebe, G. A. (1935), "Variation and correlation in grain yield among 1,500 wheat nursery plots," Journal of Agricultural Research, 50, p 331-357.

If after fitting the block model we apply our analysis to the residuals, we obtain

$$(\hat{\theta},\hat{\sigma},\hat{\beta},\hat{\alpha}) = (-.023,\ 1.0,\ .\underline{25},\ \underline{0.0})$$

and approximate confidence regions for (β,α) are shown in Figure 20.

Again comparison with Figure 14 shows that the confidence region is remarkably little changed.

7. Conclusions

From this study of real data sets, we conclude the following:

1) Very heavy-tailed distributions do occur in practice. However, they are often caused by inhomogeneity in levels and variances. Serial correlation in a stationary process makes the data more normal than the generating shocks. However, serial correlation in a finite sequence from a non-stationary process generated from normal errors produces an apparent light-tailed distribution for z_t.

2) Data collected over a long period of time usually suffer from secular inhomogeneity and would, therefore, not be expected to be normally distributed. However, much of the inhomogeneity of this kind would be eliminated or reduced when appropriate designs and blocking were adopted, and appropriate input variables were included in the model.

3) For the data we have studied which was reasonably homogeneous, the distributions are most frequently slightly heavy-tailed and/or contaminated. However, light-tailed and contaminated light-tailed distributions seem also to occur.

4) After obvious inhomogeneities in a series (for example, in most cases, due to start-up phenomena) have been allowed for, the trade off between α and β causes the confidence regions to be elongated usually and to include suitably contaminated normal distributions. We believe, therefore, that most data sets of this kind could be adequately described by contaminated normal distributions with suitably chosen α and k.

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: 2002

2. GOVT ACCESSION NO.

3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle): A STUDY OF REAL DATA

5. TYPE OF REPORT & PERIOD COVERED: Technical; Summary Report, no specific reporting period

6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s): Gina Chen, George E. P. Box

8. CONTRACT OR GRANT NUMBER(s): DAAG29-75-C-0024

9. PERFORMING ORGANIZATION NAME AND ADDRESS: Mathematics Research Center, University of Wisconsin, 610 Walnut Street, Madison, Wisconsin 53706

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: 4 - Probability, Statistics and Combinatorics

11. CONTROLLING OFFICE NAME AND ADDRESS: U. S. Army Research Office, P.O. Box 12211, Research Triangle Park, North Carolina 27709

12. REPORT DATE: October 1979

13. NUMBER OF PAGES: 67

14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)

15. SECURITY CLASS. (of this report): UNCLASSIFIED

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

MRC-TSR-2002

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

real data sets, contaminated exponential power distribution, maximum likelihood estimation, secular inhomogeneity, serial correlation

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Nine sets of real data are analyzed. The distributions within the contaminated exponential power family which best describe these data sets are obtained by maximum likelihood. It appears that heavy tailed distributions are often produced by secular inhomogeneity in mean and variance.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

221200

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)